# Report of the Research Coordination Network RCN:OceanObsNetwork

# Facilitating Open Exchange of Data and Information

May 2013

**Editors:**
**Jay Pearlman, Albert Williams III, Pauline Simpson**

**Authors (in alphabetical order):**
**J. Gallagher, J. Orcutt, P. Pissierssens, L. Raymond, P. Simpson**

**Additional Team Members**

**P. DiGiacomo, M. Kampel, T. Kawano, F. Maltz, M. McCann, B. Pirenne, I. Shepherd, and C. Waldmann**

## Acknowledgements:

The OceanObsNetwork acknowledges the support of the National Science Foundation through Grant Award No. OCE-1143683.

The Working Group members participated through Task Teams addressing various aspects of Open Data. The Teams and their members were:

**Task Team 1 on Open Data Formats and Standards**

A. Williams, J. Gallagher, C. Waldmann, F. Maltz and M. McCann

**Task Team 2 on Data Access Models**

J. Orcutt, P. DiGiacomo, T. Kawano, M. Kampel, B. Pirenne and I. Shepherd

**Task Team 3 on Data Publication/Data Citation**

L. Raymond, P. Pissierssens, and P. Simpson

These teams worked on a voluntary basis and provided the background, discussions and recommendations of this report. The editors and authors acknowledge the significant contributions of the team members.

Bibliographic Citation:

Pearlman, J., Williams, A, and Simpson, P. (eds) 2013 *Report of the Research Coordination Network RCN: OceanObsNetwork. Facilitating Open Exchange of Data and Information*. NSF/Ocean Research Coordination Network, 46pp.

## Table of Contents

# Executive Summary

**Background**

The OceanObsNetwork goals and objectives are to foster a broad, multi-disciplinary dialogue, enabling more effective use of sustained ocean observatories and observing systems. To achieve these, the activities for the RCN include a working group titled “Facilitating Open Exchange of Data and Information.” Within this area 3 task teams were created dealing with elements that impact on open exchange of data and information.

**Considerations for Open Data**

Open Data Policy is evolving with more countries adopting the approach that all data should be open and reusable with minimal restrictions. While there are technical issues being addressed for improved data discovery and access, a broad open data policy impacts the social aspects of the science research community in many ways. Both technical and social aspects are interwoven with policy. Policy and financial resources are major drivers in advancing Open Data and must be included in a comprehensive strategy and implementation of Open Data.

This report examines the foundation of Open Data and its importance to the international community, science, innovation and jobs. While the goal may be similar, the paths to Open Data are varied and drawing together a pervasive approach will take time. There are however, near term steps, technical and social, that could have significant impacts.  Stimulating interdisciplinary collaboration occurs through adoption of common standards for data exchange, creation of information brokering for improved discovery and access and working toward common or defined vocabularies. Simply finding other scientists’ data has been noted as a major barrier for research.

Open Data impinges on existing reward systems and social interactions. Areas that need to be addressed are the academic reward system (in terms of promotion and resources), the peer review panels and grant selection processes (in terms of acknowledging the importance and challenge of data collection) and the needs for acceptable citation mechanisms. Intellectual property should not be abandoned in an Open Data environment and managing IPR is necessary.

A sustainable Open Data Policy is essential and sustainability is a matter for all parties, government, private sector, academia and non-profit organizations. As full implementation of Open Data will involve a change in practices in a number of research and publication activities, an end-to-end perspective and strategy would most likely allow a long-term sustainable path to be pursued. Various business models are discussed in the paper that would not have been considered a decade ago. These range from cloud storage to publication of data with Digital Object Identifiers. These set a possible foundation for the future.

**Recommendations**

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

- Is sustainable;
- Preserves the peer review attributes of science and of publications and expands it to include data;
- Assures scientists of recognition for their 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 Intellectual Property Rights and licensing protocols;
- Is compatible with evolving national and international policies;
- Motivates participation and contributions;
- Minimizes negative impacts on existing disciplinary systems;
- Works across physical, social and economic sciences; and
- Promotes access and use by the public and policy makers.

Metrics should be established to monitor status and progress in the above areas.

*Policies should be adopted that support the sustainability of an Open Data environment.*

*Broadly inclusive collaborations across scientific disciplines need a more formal way to make data generally available. Translators (brokers) for formats should be developed as middleware.*

*Collaboration between international repositories of ocean science and other data should be encouraged both to improve efficiency and reduce costs.*

*Adoption of Digital Object Identifiers or equivalent "globally unique persistent identifiers" should be expanded and widely implemented.*

*A peer review methodology for datasets and/or data repositories should be implemented*

*Journal publishers should have a clearly stated data policy regarding supplemental material and related datasets.*

*Outreach and Capacity Building is needed for general users to be comfortable in a cross-domain data environment; such activities should be built into an Open Data initiative.*

*A Committee should be created with broad representation to make recommendations on issues, both for implementation and operations. This may be an ideal activity for a research coordination network.*

A fuller discussion of recommendations is provided in Section 8 of this paper.

# 1. Introduction

Information and technology are changing almost every aspect of our lives - how we communicate, how we work together and what we work on. It is commonly thought that if we had enough time, all information could be captured from the Internet. This may be true for some endeavors, but access to scientific data over the Internet is still not pervasive. Many datasets are not originally collected as digital artifacts and once (or if) transcribed to digital media are not necessarily available over the Internet. When data are accessible over a network, many factors impact access such as interface standards, bandwidth and discovery of what is available.

Approaches to data are evolving rapidly. There is a great deal of interest today in the idea of Open Data and the exploitation of such data for purposes ranging from the discovery of new relationships between Earth systems and climate to applications of focused advertising to specific customers (e.g. Amazon or Google). The term data used in the context of this report is broad; data are factual information used as a basis for reasoning, discussion, or calculation and is not limited to internet modalities. Open Data, as defined in the Open Data Handbook [1] is "data that can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement to attribute and sharealike." 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, 'non-commercial' 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.

"The best way to get value from data is to give it away," said EU Commissioner for the Digital Agenda, Neelie Kroes, when presenting the Open Data Strategy for Europe to public administrators. 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." [2] The Open Data strategy in Europe is part of the Digital Agenda, one of seven flagships supporting the Europe 2020 strategy to achieve growth based on research and innovation, a low carbon-economy, jobs, and poverty reduction.

In March 2012, the U.S. President provided updates to his Big Data and Open Data strategy including the details of its $200 million-plus annual big data strategy that includes lots of access to funding and data for research. "In the same way that past Federal investments in information-technology R&D led to dramatic advances in supercomputing and the creation of the Internet, the initiative we are launching today promises to transform our ability to use Big Data for scientific discovery, environmental and biomedical research, education, and national security," said Dr. John P. Holdren, Assistant to the President and Director of the White House Office of Science and Technology Policy.

"Data are motivating a profound transformation in the culture and conduct of scientific research in every field of science and engineering," NSF Director Subra Suresh said. "American scientists must rise to the challenges and seize the opportunities afforded by this new, data-driven revolution. The work we do today will lay the groundwork for new enterprises and fortify the foundations for U.S. competitiveness for decades to come." [3]

NSF's Cyberinfrastructure[1] Framework for 21st Century Science and Engineering, or "CIF21," is core to strategic efforts. CIF21 will foster the development and implementation of the national cyberinfrastructure for researchers in science and engineering to achieve a democratization of data. The first round of awards made through an NSF geosciences program called EarthCube [4], under the CIF21 framework, supports the development of community-guided cyberinfrastructure to integrate big data across geosciences and ultimately change how geosciences research is conducted.

In summer 2012, the Japanese government announced a new Open Data strategy, with the intention of connecting the country's governmental, industrial, and academic sectors. Japan is set to have a record year in 2013 for Open Data projects, with open government advocates leading the way. [5]. On March 1 2013, the Japanese Government announced that it will launch a National Open Data Portal as part of its commitment to create an environment where data are used by citizens to promote innovation, creative industries and knowledge-based services.

The Government of Australia supports the concept of mash-ups of Open Data through its government 2.0 task force. "There are now a growing number of these 'smashup' initiatives creating new services, often with simple tools and the energy and ingenuity of people and communities keen to solve a problem or create an opportunity. We want you to show us what you can do with better access to re-usable public data and plenty of imagination by creating mashups using Australia Government data." [6]

[1] The term cyberinfrastructure came into significant use in the Atkins Report to NSF, "Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. 2003. It's definition was complex and thus there have been a number of simplified alternatives, one of which is "Cyberinfrastructure consists of computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked together by software and high performance networks to improve research productivity and enable breakthroughs not otherwise possible." A more comprehensive discussion is given in Stewart, C, et. al. "What is Cyberinfrastructure?" SIGUCCS'10, October 24–27, 2010, Norfolk, Virginia, USA.

Looking at the science community, Borgman [7] lists four reasons for sharing data in the sciences:

1. Reproduce or verify research;
2. Make results of publicly funded research available to the public;
3. Enable others to ask new questions of extant data; and
4. Advance and accelerate the state of research and innovation.

The identification of Open Data is an issue that is recognized globally. The EU Open Data strategy asserts that opening up access and reuse to public sector data offers major opportunities not only for innovation and growth and for more informed science, greater public participation, but also for addressing societal and environmental challenges. However, the implementation of an Open Data environment has its own challenges and the science community has large variations in the openness of its data.

From a broader perspective, the emerging modern deluge of data in both government and science [7] stands in stark contrast to the lack of progress of sharing both within and across agencies and disciplines. The cost of this disconnect is brought to the fore by EU estimates of more that $50B a year in economic benefits that could be derived [2]. The recent lessons learned in what has come to be termed ClimateGate include the need to make data much more available to researchers as well as the public in general. The UK Joint Information Systems Committee (JISC), and the Program Manager, Simon Hodson, noted: “Climate science is by no means unique in the need for researchers to analyze complex data from a number of different sources. The aim of this investment is to improve the way research data are managed in UK universities. By showing how it can be made more open, these projects will help achieve proper recognition for the essential place of data creation and management in the research process.” [8] Had the climate data been more open and available, much of the controversy could have been avoided.

## 2. Challenges for an Open Data Environment

Acknowledging the current environment of data sharing, identification of the challenges and the opportunities for expansion of sharing provide the basis for recommendations of this report. The discussion starts with an identification of issues for an Open Data policy.

The first step is to converge on a definition of “Open Data”. There is general consensus that “Open Data are data that can be freely used, reused and redistributed by anyone - subject only, at most, to the requirement to attribute and share alike.” [9]. The philosophy behind Open Data has been a tradition of science, but the term “Open Data” itself is relatively new. 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 Open Data in science, moving forward to a uniform approach 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. quality control, 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/knowledge through models ranges 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.

As long as the research scientist who acquires data analyzes them, draws conclusions about processes the observations were designed to reveal and reports these conclusions, the problem of obscure or idiosyncratic data formats is hidden. When it is necessary for others to review and reanalyze these data or the original researcher needs to recall later what formats the data conformed to, standards for data format become important. When data are open, placed in repositories where anyone may access them and reanalyze them, Open Data formats and standards are absolutely necessary. Even with the creation of universal standards, research scientists must individually or collectively adopt and implement these. A survey of repositories shows considerable differences in terms of openness, which draws us to the conclusion that the Open Data community acts in a very uncoordinated way at the moment and needs to be aligned in the future. To move forward, the broad community must understand the benefits of Open Data in justifying significant investments in data formats and interfaces necessary for Open Data.

Open Data Exchange faces many difficulties with access models, publications and citation, and with formats and standards that operate at the most basic level. Considering only digital formats for data (not necessarily a complete set), the file format and Internet web services are the standards that apply. If the data are not digital, there are almost certainly supporting data describing physical samples such as biological samples, geological cores, and photographs or other non-digitized records that do have a digital existence. These then must be readable, searchable, downloadable, and informative across platforms.

Looking beyond observations to the sharing of data, file transfers make up most of the networked data access. For users this means they must first find the files they want, transfer them and then figure out how to read them. The challenge is that the data comes in arbitrary size packets and the user must break them apart and find the data in which they are interested. In the course of this process, it is very likely that the user will transfer and process far more data than they actually need or use. In a survey and analysis of open-data repositories (K. Braunschweig, et al) [10], two major problems common to almost all platforms were dead links and a plethora of different file formats. In contrast, 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.

Ultimately, we come back to the intent of Open Data – not only that it is available at minimal cost, but also that it is accessible and easy to use in a variety of contexts. Critical to any data system is thus the concept of interoperability. It may seem that this means simplicity in transferring and rendering useful data and information across systems. But there is significantly more to 'interoperability.' 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 [11]. The second approach is to mandate certain standards that must be followed by each domain system so that the different systems will be interoperable [12]. 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 largely 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, their competitors could gain an "unfair advantage".

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 [13]. Only recently has there been provision through the use of Digital Object Identifiers to enable effective referencing of data sets. In a study of linking data to publications, a project by T.W. Pace was done to help researchers link their datasets to their publications, thus creating "enhanced publications". Even when scientists do want to make their supplementary research material available, for example software and mathematical proofs, but they may need assistance in doing so. This is in part "ease of use" and part motivation and priorities [14]. Motivation is also needed to bring "dark data"[2] into the open, making it web accessible.

As an additional human factor, in some disciplines, there is no documentation of accepted

[2] Digital data not accessible using the web.

practices (sometimes called "best" practices) that can be referenced by people in other disciplines. Even within a discipline such as oceanography, there are subcultures that guide best practices and approaches to data release, formats, languages or semantics, quality assessments and communication protocols. For effective interoperability, there must be some form of translation between domains – between physical, biological, biogeochemical, social research scientist and many other fields. This has been a challenge in addressing multi-disciplinary issues such as climate change, fisheries, food and water resources. The challenge is to make Open Data usable for research across domains.

Recognizing these and other challenges, Nielsen [15] 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:

> *I believe 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.*

## 3.0 Perspectives on Open Data: Examples

One means to understand approaches and motivation for Open Data in a variety of environments is to examine use cases. With a subject as broad as Open Data, the use cases can be quite diverse. In his study of Open Data and science research mentioned earlier, Nielson [15] has provided a simple example of Open Data and its impact on research; the example is of the use of open access for solving a mathematical problem. Tim Gowers, a young mathematician at Cambridge, had a problem he wanted to solve, but had run into a variety of barriers. He exposed the problem to the public in the *Polymath Project.* The web site he established, polymathprojects.org, provided an environment where ideas could be quickly improved by a variety of people. The *conversation* was scaled up beyond the point where normally an individual would rely upon fortuitous serendipity, but where the large numbers of mathematicians were able to manage insights as a matter of course. The participants included a Field medalist, a professor of mathematics at UBC and a high school math teacher. The problem that Gowers wanted addressed was solved in only 37 days although it involved 27 people, 800 comments on formulae and 170,000 words. This example of crowd sourcing has, at its core, the availability and use of Open Data.

An example of a similar approach to discovery involving other data sets is the Galaxy Zoo (www.galaxyzoo.org), which recruited more than 200,000 online volunteers to help astronomers classify galaxy images. These were photographs taken automatically by a robotic telescope and the images were really studied for the first time by the participants. The online community uncovered a new form of galaxy, in addition to the usual spiral or elliptical galaxies.

**The Human Genome Example**

The Bermuda Agreement [16] in 1996 changed the complexion of genetics research fundamentally, taking the related research community from a limited practice of openness to one in which the Human Genome Project shared the human genome broadly. The previous cautious approach to opening these new data gave way to the public publishing of the genome because the attendees could envision the enormous benefit of a common commitment to publishing and on-line data access. The relevant funding agencies quickly required that all scientists working on the human genome make the data openly available. Perhaps surprisingly, genomic data associated with other life forms continue to be held privately in spite of the great potential value to society. This echoes scientific attitudes more broadly in that, lacking incentives to make data open, scientists will tend to hoard their data with the consequence that many important observations will be lost to the scientific enterprise. Considerable leadership was needed to open the human genome for unlimited study by all.

**The Ocean Observing Example**

Looking at the US National Science Foundation (NSF) programs, Nielson also used the NSF Ocean Observatories Initiative (OOI) as an example of Open Data in the environmental sciences. He stated:

> *In September of 2009, an organization called the Ocean Observatories Initiative began building a high-speed network for data and electricity on the floor of the Pacific Ocean. They're extending the Internet to the ocean floor, with the eventual plan being to lay 1,200 kilometers (750 miles) of cable, from the shores of Oregon all the way up to British Columbia. This underwater Internet will range to more than 100 kilometers (60 miles) offshore. When it's complete, all manner of devices will be plugged into the network, from cameras to robot vehicles to genome-sequencing equipment. Imagine a volcano erupting underwater, and nearby genome-sequencing equipment switching on to take genetic snapshots of never-before-seen microbes vented during the eruption. Or imagine a network of thermometers and other sensors mapping out the underwater climate, in much the same way the Sloan Digital Sky Survey (SDSS) is mapping out the universe. But the Ocean Observatories Initiative is going even further than the SDSS, making their data openly available right from the start, so anyone in the world can immediately download the data, looking for new patterns and asking new questions.*

In Canada, the NEPTUNE and VENUS projects of Ocean Networks Canada have been in operation since 2009 and 2006 respectively. The systems extend the Internet underwater up to 300 km off the coast of British Columbia and are hosts to over 100 instruments. Those networks collect on the order of 100 millions of individual scalar measurements every day, as well as currents, acoustic and video data. All the data are made public immediately, most in quasi real-time.

These examples illustrate the potential of open observatory data. They do not mention the corresponding issues that come with broad use of open data – quality control, interfacing with users to answer questions, notifications of updated data as recalibration occurs, intellectual property considerations, etc. Some of these will be addressed later in this paper.

### US National Science Foundation Example

The US National Science Foundation is working in a direction that seeks to make data resulting from the agency's support more open and better managed. The specific policy is:

> *Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. Privileged or confidential information should be released only in a form that protects the privacy of individuals and subjects involved. General adjustments and, where essential, exceptions to this sharing expectation may be specified by the funding NSF Program or Division/Office for a particular field or discipline to safeguard the rights of individuals and subjects, the validity of results, or the integrity of collections or to accommodate the legitimate interest of investigators. A grantee or investigator also may request a particular adjustment or exception from the cognizant NSF Program Officer.*

There is no time constraint on providing access to the data. The NSF Ocean Sciences has a two-year requirement that imposes a deadline (effective May, 2011):

> *Principal Investigators are required to submit, at no more than incremental cost and within a reasonable time frame (but* ***no later than two (2) years after the data are collected),*** *the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF/OCE grants to the appropriate Data Center (See appendices below or consult with the cognizant NSF Program Officer).*

Generally, data submission policies have not been evenly enforced by NSF program managers (e.g. Borgman, 2012) and many Principal Investigators delayed release beyond the two-year limit or sought and received exceptions.

In the recent past, the NSF has asked that all proposals include a data management plan:

> ***Plans for data management and sharing of the products of research***
> *Proposals must include a supplementary document of no more than two pages labeled "Data Management Plan". This supplement should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results (see AAG Chapter VI.D.4), and may include:*
>
> 1. *The types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;*

2. *The Standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);*
3. *Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;*
4. *Policies and provisions for re-use, re-distribution, and the production of derivatives; and*
5. *Plans for archiving data, samples, and other research products, and for preservation of access to them.*

The NSF is moving forward with a process, which will capture more of the data collected and created through their sponsored programs. It is anticipated that requirements will continue to evolve toward increased specificity in terms of requirements in coming years.

There are issues that the above policies of NSF do not effectively address. Once a grant is completed, who is responsible for the long-term sustainability of the data storage and access? Who is responsible for the evolving interoperability of such data across domains? Who is responsible for updating formats as obsolescence occurs? Who maintains the data provenance and corrects errors discovered by scientists? In essence, these questions raise issues about the both the technical and human factors side of Open Data and the sustainability of the Open Data paradigm. Providing data, as the above NSF directives address, is only one element of an effective Open Data program. To have an effective Open Data policy, a number of core issues must be addressed which go beyond just the provision of initial data access.

## 4.0 Core Issues for Open Data

From the use cases above and others, it is seen that Open Data has different implementation approaches. However, the overall definition of Open Data has continued to converge over the last decade. The definition from the Open Data Handbook is: “Open Data is data that can be freely used, reused and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike.” This definition appears to be generally accepted. Opendefinition.org expands the definition, addressing both the technical and social aspects of Open Data and its use. [17]

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 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; and
3. Sustainability for future access.

There is overlap in the above issues and the boundaries between them 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.

Interoperability is sometimes thought of as transferring and rendering data and information across software systems. This is often seen as a simplified network diagram connecting various users and suppliers. However, Figure 1 (Palfrey and Gasser, 2012 [18]) illustrates a broader view with the technological layer at the bottom and the institutional layer at the top. This approach is gaining wider acceptance in the discussions of broad interoperability in a multi-disciplinary research environment.

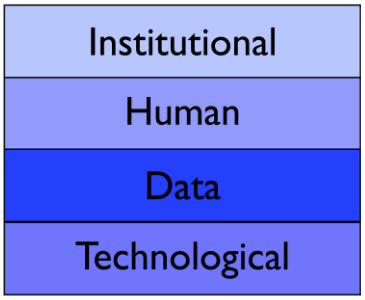


Figure 1. A layered view of interoperability according to Palfrey and Gasser [18].

The first layer, technology, is analogous to the train tracks in transportation and the switches and routers in networks. The layer assumes the ability to connect between systems through explicit interfaces.

The second layer nearly parallels the technological layer; the receiving party must be able to understand the data received.  If understanding is not possible, neither the data nor the technological layers are useful. A good, simple analogy is receiving an email, clicking on an enclosure, and finding it can't be read by any software resident on your computer.  This was once quite common when sending information from a Mac to a PC or the reverse. In spite of the possible desires of Apple or Microsoft to "lock in" their products, interoperability has become quite common for this case and serves the interests of the vast number of users.

The third layer of interoperability is the human layer. Assuming that the mail enclosure problem above has been solved, it's still important that the humans involved are able to understand what has been passed in the lower two levels. Generally, a common language (e.g. English) is needed, but that's a rudimentary beginning. Interoperability does require a strong commitment by people to working together as humans. This is often hard work, may require learning a number of new skills, and takes time away from other tasks. Without this willingness, however, interoperability is impossible. In the analogy, the email attachment is opened, it is in the correct language and it is then put into a folder for reading later, never to be viewed again. A step further is when the material is skimmed and has a few incorrect statements – which then causes the rest of the material to be disbelieved. The matter of trust is an important part of the human level. Part of the trust is built through understanding provenance, or the history of the data and information. Part of the trust comes from understanding the uncertainty in the data, or "data quality". More broadly, information on the data (metadata) covering the technical and human aspects must be documented somewhere accessible to be useful. Integrating and working the social and human component is perhaps the greatest challenge posed by the construction of the NSF Ocean Observatories Initiative as well as VENUS/NEPTUNE and others.

The highest level, Institutional, societal systems must also interoperate. The legal aspects (e.g. The treatment of intellectual property rights) are often quite important and some examples are given in the next section. This doesn't imply, for example, that two countries must have the same laws, but that the laws in each country provide enough commonality for the interests of each party to be protected. In an analogy, the wall plugs in each country may be different, but their configurations are know and translators (wall plug adaptors) can be built that reliability address the electrical access needs of travelers.

Table 1: Interoperability matrix status for four observatories (US & Canada). The color code is associated the traffic lights – go, caution and stop or in the case well-developed, partly complete or none/elementary. Reds indicate substantial problems representing largely the human and institutional aspects of interoperability.

| | IOOS | OOI | NEPTUNE and VENUS | NDBC Radar |
|---|---|---|---|---|
| Institutional | 🟥 | 🟥 | 🟩 | 🟥 |
| Human | 🟥 | 🟥 | 🟨 | 🟨 |
| Data | 🟨 | 🟩 | 🟩 | 🟩 |
| Technological | 🟨 | 🟩 | 🟩 | 🟩 |
| | | | | |

Long-term interoperability can supported when each of the core issues are addressed. While the three issues are interlinked, a discussion of each can highlight some of the challenges in each area.

## 4.1 Issues: Discovering and Accessing Data

Open Data Exchange faces many difficulties with access models, formats and standards that operate at the most basic level. Considering only digital formats for data (not necessarily a complete set), the file format and Internet web services are the standards that apply. If the data are not digital, there are almost certainly supporting data describing physical samples such as biological samples, geological cores, and photographs or other non-digitized records that do have a digital existence. These then must be readable, searchable, downloadable, and

informative across platforms and online. The Internet data that are stored may be discovered with DAP or a similar application. Access requires greater user sophistication, and Outreach or Capacity Building may train users to seek support services to enable them to access otherwise unfamiliar data formats. Translation middleware between one format and a better supported one may become an attractive commercial market and thus be developed.
Open Data are accessed using a variety of tools, some of which depend on a lower layer of computer networking technology and some that do not. For example, data held in a spreadsheet and passed between researches on a USB memory stick fits the definition of open data. However, for many, the exposure provided by the Internet for open data is so much greater than person-to-person transfer that it effectively alters the process in a qualitative way.

Access to open data using the Internet can be broken down into two modes: 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 users 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 FTP or HTTP, although this short list is not exclusive. Both FTP and HTTP 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 used by the Web, has the advantage that it’s widely supported, every host on the Web has a HTTP server running. Both protocols support both anonymous and authenticated access as well as logging all accesses. Beyond these features, file transport is a very simple paradigm that is central to most computing machinery (although this is somewhat hidden by some mobile computing tools like smart phones).

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 ‘fit together’ 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, different return types and 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, THREDDS/DAP developed by Unidata and OPeNDAP, et cetera.[3] 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 because users do not have to decode files, they 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.

However, while sophisticated data access protocols provide tangible benefits for both sides of the data access equation, there is evidence that they are used less frequently than file access protocols. The American Consumer Satisfaction Index (ACSI) conducted a survey that included questions about access to NASA's online satellite data and found that less than half of data requests used advanced protocols like WCS or DAP. Instead most users relied on FTP or HTTP to transfer files.

*A recommendation of this report is to determine if the dominance of file transfer relative to web service access is isolated to NASA's data holdings (it might be specific to satellite data, which have some unique characteristics relative to in situ data) or if it is widespread. If there are impediments to using tools like DAP or WCS, those need to be understood.*

In addition to information about protocols used to access data, the ACSI survey also contained in formation about issues that straddle the boundary between data access and data quality. The survey found that:

- Users had problems finding data they knew were accessible;
- Users had problems finding documentation for datasets;
- Documentation quality was often poor;
- Network reliability was often poor and users would benefit from more robust protocols (of any sort);
- Data formats were not appropriate for more users; and
- Users were not aware of technical support services that were available.

For example, data should be bound more tightly to their documentation. One potential solution is to include in individual data files a reference to the associated documentation. This is effectively the approach taken with XML when a specific XML file contains a reference to its schema. This could be realized fairly easily by embedding a URL in a file such as a netCDF or HDF file as the value of an attribute. Here, DOIs might also play an important role in moving forward solutions to these issues.

[3] WCS, WMS, DAP and similar protocols are often called 'Web Services' because they provide remote computing capabilities while building on the Web software stack and infrastructure.

## 4.2 Issues: Data Quality and Fitness for Purpose

Diversity of data is increasing. 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, biofouling or uncertainty due to a sampling process. When data from one 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 contributing data sets. This is a general issue and is very 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 questionable data for failing 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. Quality information is valuable for certain kinds of data and the absence of qualifiers make some data worthless [19]. Various international projects are looking at quality indicators. CEOS QA4EO [20, 21] is a GEOSS quality assurance protocol. GEOVIQUA [22] in Europe has focused on adding rigorous quality specifications to the Global Earth Observation System of Systems (GEOSS) spatial data in order to improve reliability in 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. This is addressed by QARTOD [23].

Quality and other factors are tied to the ability to use the data for research and other purposes. Criteria in this area concern completeness, correctness and consistency of the data for such use and is applied particularly when data use crosses domain boundaries where interpretation may be made in the context of the local domain culture. Provenance and traceability support knowledge of uncertainty in the data and are an important element in user understanding what may be the fitness for purpose of the 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” [24]. This is a particular problem in long term studies such as climate where the data are produced by a multitude of sensors that may not be and sometime cannot 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.”

## 4.3 Issues: 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, involves a combination of resources, human factors and policy. These areas need to address consistently over a long period of time to motivate a cultural change that may take a generation. What are the issues? As noted above, data must be discoverable, accessible and of "known quality" as a first step. Ownership of the data and the innovations fostered is imbedded in intellectual property rights (IPR) that govern who benefits. As will be noted later, the laws regarding ownership of the outcomes of scientific research in the US changed to allow universities to retain the IPR. This became a significant business opportunity for educational institutions. Publishing houses, both profit and non-profit including science and technical organizations (IEEE, AGU, AAAS, etc) 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. For the Principal Investigator, promotion was based on high quality publications of research. Data was used in analysis for such publications and was not released, if at all, until the research was published. 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 with 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? In essence, there are challenges as how to define and apply open data.

The National Science Board initiated a study in this subject in 2010, raising the above questions and many more relating to the management, business models and rights with respect to Open Data [25]. 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."

While the core technical capabilities exist for handling 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 evolving their own policies with potentially important variations in the implementation details. Thus, leadership in implementation approaches from government is a critical step in providing sustainability of Open Data.

## 5.0 Uses Not Intended – the need for 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 application that envisioned – uses that were not the intent of the 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 4 above. In recent years, important programs and initiatives are focusing on this challenge, including: (a) the European Infrastructure for Spatial Information in the European Community (INSPIRE) [26], and the Global Monitoring for Environment and Security (GMES) [27]; (b) the US National Spatial Data Infrastructure (NSDI) [28], Data Observation Network for Earth (DataOne) [29] and the recent EarthCube [4]; (c) the international initiatives Global Earth Observation System of Systems (GEOSS) [30].

There are several well-known disciplinary infrastructures, such as: WMO Information system (WIS) [31], the Global Biodiversity Information Facility (GBIF) [32], the Pan-European Infrastructure for Ocean & Marine Data Management (SeaDataNet) [33], the US CUAHSI Hydrologic Information System (HIS) [34], the IODE infrastructure for oceanographic data and information exchange [35], and a global geology information network, OneGeology [36]. There are others under development, including: the European Plate Observing System (EPOS) [37] and the GEO Biodiversity Observation Network ((GEO BON) [38].

According to a study of the European Commission, [39] 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 modeling 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. This goes beyond the technical interoperability addressing conceptual/semantic models interoperability. Figure 2 shows the current version of LCIM [40].

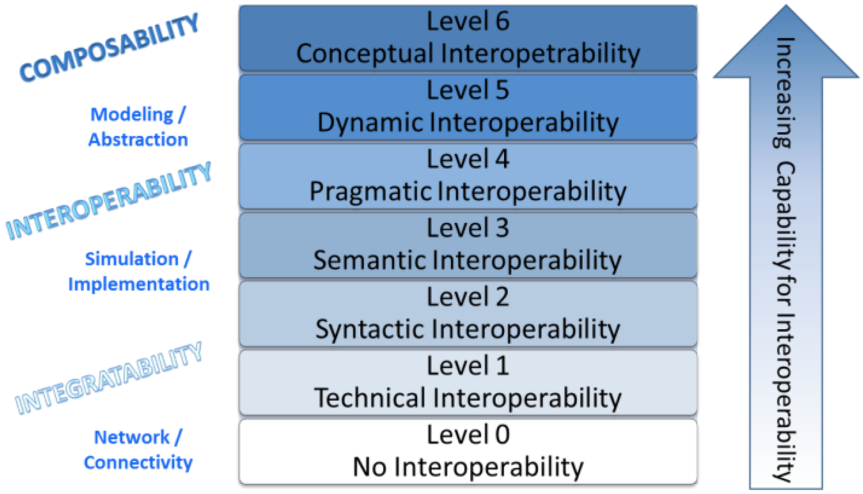

COMPOSABILITY
Modeling / Abstraction
INTEROPERABILITY
Simulation / Implementation
INTEGRATABILITY
Network / Connectivity
Level 6
Conceptual Interopetrability
Level 5
Dynamic Interoperability
Level 4
Pragmatic Interoperability
Level 3
Semantic Interoperability
Level 2
Syntactic Interoperability
Level 1
Technical Interoperability
Level 0
No Interoperability
Increasing Capability for Interoperability

Figure 2. Levels of conceptual interoperability [40].

These levels are a finer gradation of the technical levels in Figure 1.The seven layers of the LCIM are as follows:

- Level 0 (No Interoperability):
  Stand-alone systems- no data is 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 its use are “understood” by the participating systems, and the context in which it is 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 as “fully specified but implementation independent models” as suggested in Davis and Anderson, enabling their interpretation and evaluation by other engineers.

Convergence is a challenge in moving interoperability to higher levels. Care must be taken to define these objectives for this. In an analogy, electric plugs are standardized within countries or within regions, yet the standards are not widespread so that travelers cannot travel around the world and use their electrical devices without forethought. The solution has been a converter, a device that goes between the wall and your electric unit to make the two “compatible”. In the software world such a converter or mediator can be middleware that supports translation from one community’s formats and standards to another community. The cost of the wall socket converters is small compared with the conversion of all electrical systems. For software interoperability, the cost argument is not so straightforward. However, 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 disciplinary 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 along with different vocabularies (these may be called cultural aspects). Bridging across disciplines is a complex challenge. A brokering approach was recently introduced to handle such differences without limiting the autonomy and without putting a significant investment burden on existing disciplinary systems [41]. It is noteworthy that there exist different interoperability levels –as depicted in Figure 1. Brokering technology can facilitate interoperability at all levels shown in Figure 2; the brokering approach integrates and supplements the standardization approach building effective systems of (autonomous) systems. Ultimately, interoperability solutions of a global nature will be a combination of middleware, standards and a compendium of best practices.

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. 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. Examples of these kinds of standards are TCP/IP for computer networking and ISO 19115 for geographic metadata. Augmenting these formal standards are community standards that, while they lack formal ratification by an established standards body, are in such widespread use as to be de facto standards. Examples of these include HDF4 and HDF5, which are used by NASA to store much of their satellite data and GeoTIFF which is widely used by the Geographic Information Systems (GIS) community to store and transfer data. Thus, by combining networking (TCP/IP), data format (HDF5) and metadata (ISO 19115) standards, a modest level of technical interoperability can be achieved.

## 6. Governance – Business Models and Policy

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 to create related information in a manner where financial, career growth, grant selection or other rewards are available to participants. 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 [42] has contrasted several observational programs and her approaches can be summarized as in Figure 1.

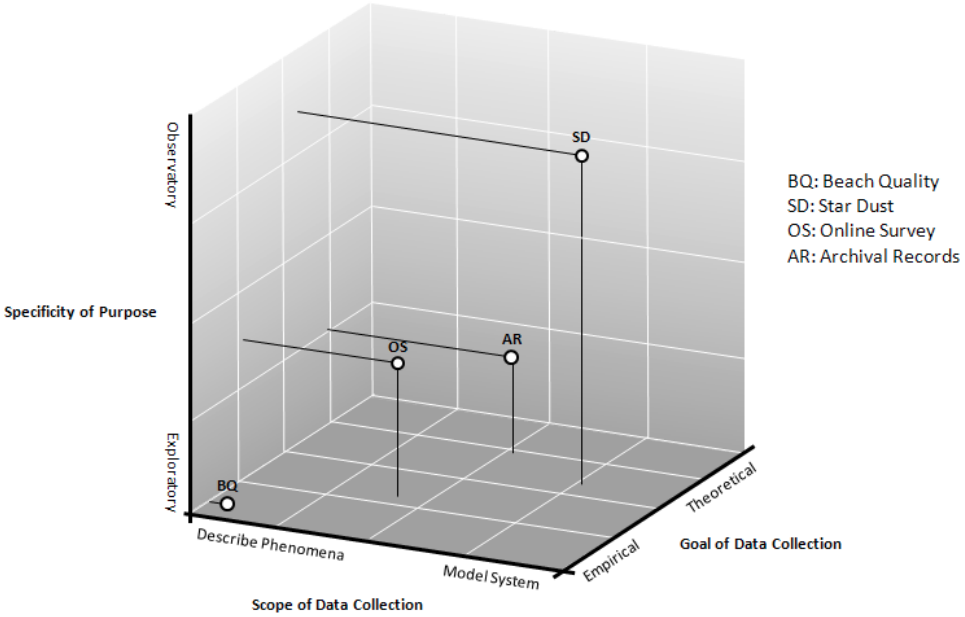


Figure 3. The cube shows the diversity of data collection and analysis efforts that should be addressed under Open Data policies [42]

One of these, 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. In Figure 3, beach quality tends to lie to the left in the scale of describing phenomena/modeling while stardust lies to the right.

For the beach quality measurements, the three classifications lie toward the left:

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 on the scale from *By Hand* to *By Machine*.

For Star Dust there are contrasting descriptions where the classification falls to the right:

Specificity of purpose: Observatory
Scope of Data Collection: Model system
Approach 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 becomes 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.

## 6.1 Business models

Two decades ago, the primary business model for data and scientific information comprised of subscription services. Whether the subscription was paid by a University library on behalf of the students or by an individual for a scientific journal, a fee structure would carry most of the cost of operations. For scientific papers, there were voluntary page charges for standard length papers and mandatory fees for longer papers.

Over the last decade, the boundaries of the subscriber model have changed. With the advent of inexpensive storage and pervasiveness of high speed Internet, traditional paper-based access to information and sales is transitioning, forcing the consideration of alternative business models. The models are evolving rapidly and the environment is competitive. Examples of alternatives are:

1. Amazon built an array of servers to support their online business. They now offer space on the servers “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 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 supported, for science and engineering, by federal and state governments. 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 are identified and potentially accessible. The DOI are also a tool for locating *or discovering* papers. This is discussed in detail in Section 8 below.
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 (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 [43], 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) [44] 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) [45] 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[4].

6. 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 more 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 impacts 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, science 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.

[4] For example, the annual fees for various types of organizations are:

| | |
|---|---|
| US/EU University | $500 |
| US/EU small company (<$2,000,000) | $2,200 |
| US/EU Technical Committee Member | $11,000 |
| US/EU Principal Member | $55,000 |
| Djibouti University | $165 |

## 6.2 Licensing Options and Policy

**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 (Bayh-Dole Act, 2012). 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. 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. Colloquially, this created the idea of "follow the money." The entity supported by funding from the federal government holds the intellectual property rights for work done by that entity. Subcontracts, for example, 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.

**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) (BSD Licenses, 2012) 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) will likely use BSD to license the Open Data (and open software) available through the OOI. 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 Free Software Foundation (Free Software Foundation, 2012), which states that it is compatible with the GNU General Public License (GPL) (GNU General Purpose License, 2012).

> The two clauses in BSD are:
>
> Copyright (c) <YEAR>, <OWNER>
> All rights reserved.
>
> Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:
>
> 1. Redistributions of source code must retain the above copyright notice, this

list of conditions and the following disclaimer.

2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

*THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.*

The copyright could be, for example *Copyright © 2012, Ocean Observatories Initiative, Cyberinfrastructure Implementing Organization, UC San Diego* for software or data from the Ocean Observatories Initiative. This copyright statement would have to be repeated in redistributions of software or data. In addition, the second statement must also be included and is a “hold harmless” clause. That is, the use of the data (or software) indemnifies the provider (e.g. OOI CI) against any future damage resulting from the use of licensed software or data.

**GEOSS**

GEOSS has integrated a legacy approach on licensing of Earth observation data and information into a summary white paper [46] for the global observatory community entitled *Legal options for the exchange of data through the GEOSS Data-CORE*. The white paper list four principles:

1. The data are free of restrictions on reuse;
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.

2), 3), and 4) are permitted, but not required. 1) is the most important and declares that data reuse is unlimited. 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) protection 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. (Summary White Paper, 2011)*

Much of the discussion in this section deals with the US in which intellectual property, copyright and patents are federal government functions. This is not the case in Europe in which it is important not only to comply with EU law, but with local state law as well. Creative Commons [47] provides a means for dealing with the multiplicity of laws.

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. Another alternative to this is Creative Commons (Creative Commons, 2012).

**Creative Commons**

Creative Commons was included in the GEOSS *Summary White Paper* [46] and details are available at *Creative Commons* [47]. Palfrey & Gasser [18] support the idea of using this approach to managing intellectual property by taking a permissive approach for exchanging works across systems, applications and components. There are six licenses available, extending from a rights management system much like BSD to much more constrained systems, which prevent modification of the software or data or its use for commercial purposes. There are substantial costs in developing a new system from scratch; creative commons lowers this transaction or legal cost to a very low level. Creative Commons [34]) even provides a software tool for choosing the appropriate license in response to questions about constraints. CC urges users to exploit the licenses to reduce legal costs and emphasize creativity on the part of the owner of the intellectual property. There are license catalogs for many countries, including members of the EU that recognize the national constraints.

The CC licenses comprise three layers (Creative Commons [47]) as extracted from CC information:

**Machine Readable**
**Human Readable**
**Legal Code**

The Creative Commons public copyright licenses incorporate a unique and innovative “three-layer” design. Each license begins as a traditional legal tool, in the kind of language and text formats that most lawyers know and love. This 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 format that normal people can read — the Commons Deed (also known as the “human readable” version of the license). The Commons Deed is a handy reference for licensors and licensees, summarizing and expressing some of the most important terms and conditions. Think of the Commons Deed as a user-friendly interface to the Legal Code beneath, although 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 — 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 CC Rights Expression Language (CC REL) to accomplish this.

The ability to provide *open and easy discovery and access is a key attribute of whatever licensing modality is adopted. Practically, users should be unencumbered, as much as practical, by traditional barriers such as specific/unique data formats and isolated "silo" datasets. The evolution in licensing is supporting the directions and creative commons is providing a foundation for open exchange. However, the issues associated with data citation are still present and recognition, particularly for career advancement, still needs to be addressed and accepted by the community.*

## 7. Data Publication/Data Citation

Can data publication/data citation offer a solution to some of the human motivation issues discussed in Section 2 of this paper? If so, how would it best be implemented? There has been much discussion in the community (D.S. Sayogo [13]) on incentives for researchers to share their data as Open Data i.e. information that can be freely used, re-used and redistributed by anyone - either free or at marginal cost. In this paper we define Data Publication as making data freely and as permanently available on the Internet along with easily digestible information as to its trustworthiness, reliability, format and content to enable discovery and re-use. Data Publication can take the form of: Standalone Data Publication; Data Publication by Proxy; Appendix Data; Journal Driven Data Archival; Overlay Publication [48]).

In the past, there was little incentive for a researcher to make data available. The data are needed to prepare a peer reviewed research paper, which is essential to further a research career (publish or perish). The effort to produce data is often not highly rated yet research progress is slowed without data being available to the research community as a whole.

It has been observed 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. 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 the frequent lack of the most basic back-up practices, makes data unavailable and constantly at risk of being lost. Adding 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 Task Team on data access models).

Often there are insufficient incentives for data submission by researchers, resulting in low submission rates and even when submitted, they lack a bare minimum of metadata. The problem is in part cultural. Publications of data do 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 challenging environments.

In the research community, peer review is the accepted process for evaluating the scientific quality of scientific work. Community acceptance of an agreed peer review procedure for data publication is not currently available and is due; yet it is essential to support reliable and trustworthy data sharing and to offer data creators the kudos for promotion. Data publication in data repositories ensures provenance, permanence, attribution and quality of format and metadata; at present they do not guarantee the scientific quality of published data (which requires domain experts). Emerging open access data journals that publish papers on the management of data and articles on original research data(sets) (e.g., *CODATA Data Science Journal*) are now offering peer review (e.g., *Earth System Science Data* and *Geoscience Data Journal)*. Alongside data journals, data repositories offer Open Data publication.

Data publication (making the data sets available to users) with supporting data citation (referencing the data sets) 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 the use of data citation metrics by institutional management as an element in performance assessment and career advancement decisions.

A successful example of motivating data sharing is the work of 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) have developed and executed a pilot project (see http://www.iode.org/datapublication) related to two Use Cases:

1. Data related to traditional journal articles are assigned persistent identifiers referred to in the articles and stored in institutional repositories;
2. Data held by data centers are packaged and served in formats that can be cited: The Published Data Library (PDL) and Published Ocean Data repository (POD).

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 (DOIs), enables accurate data citation. The project will also be investigating URI’s and NameIDs. Associated new data repositories are meant to be complementary to national (e.g. IODE NODCs, ICSU WDS) and thematic data centers, rather than a replacement. A "cookbook" will be published that provides extensive instructions and guidelines to scientists as well as the data publication process to repository managers.

The advent of Funding Agency mandates for Open Data, such as the requirement in the NSF Data Management Plan and the European Commission's recent recommendation for open access to scientific data is expected to stimulate authors to make data available. Data Repositories and data journals providing citation metrics will offer evidence of compliance. In addition to standard search engines, new secondary services like the Thomson Reuters *Data Citation Index*, will facilitate discovery, use and attribution of datasets and data studies by connecting researchers and data repositories around the world.

## 7.1 Use Cases

**Use Case 1:**
**The MBL WHOI Library** https://darchive.mblwhoilibrary.org/ 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 we have retroactively linked data to articles after publication. The system has been in operation for almost three years, with interest developing in the last six months, as community discussion increases and authors become aware. 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"; "Happy."*

Scientists are now becoming aware that DOIs offer the means to easily cite their datasets and gain important citation metrics. Before DOI, researchers found problems in knowing how to cite their datasets and were often not required to do so by publishers. Library use of DOIs was at the forefront, but they are becoming one of the de facto standards for data citation within data repositories and are being facilitated in such services as NASA's EOSDIS, Pangaea etc.

Publishers are now joining the Linked Data community acknowledging the importance of datasets supporting and within published articles e.g. Nature Publishing developed a platform in 2012. 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 field, publishers use PubMed and in fact are required to do so by such Funding Agencies as the National Institute of Health and the Wellcome Trust. Many publishers do not yet have an identifiable policy dealing with supporting datasets [49].

Because of the assignment of DOIs, Elsevier Publishing sought collaboration with the MBLWHOI Library. Article records in ScienceDirect now contain links to datasets deposited in the Woods

Hole Open Access Server (WHOAS) 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. The WHOAS system of linking data to the articles in ScienceDirect website was implemented in May 2012. Ways of exposing linked data is a concept that is still emerging –Tim Berners-Lee's vision [50] is to build a web for open, linked data that could do for numbers what the Web did for words, pictures, video; unlock our data and reframe the way we use it together.

Another outcome of the project includes tools and procedures developed by the MBLWHOI Library and the Biological and Chemical Oceanography Data Management Office (BCO-DMO) that automate the ingestion of metadata from BCO-DMO for deposit, with a copy of each dataset into the Institutional Repository (IR) Woods Hole Open Access Server (WHOAS). 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 utilizes 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 when she presents her work, the most asked about functionality is the DOI and the ability to cite the data. This collaboration has been in place for approximately two years.

**Use Case 2:**
**The Published Data Library (PDL)** is implemented by the British Oceanographic Data Centre (see https://www.bodc.ac.uk/data/published_data_library/). 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 NERC's Environmental Data Centers, the repository assigns DOIs obtained from the British Library/DataCite to appropriate datasets.

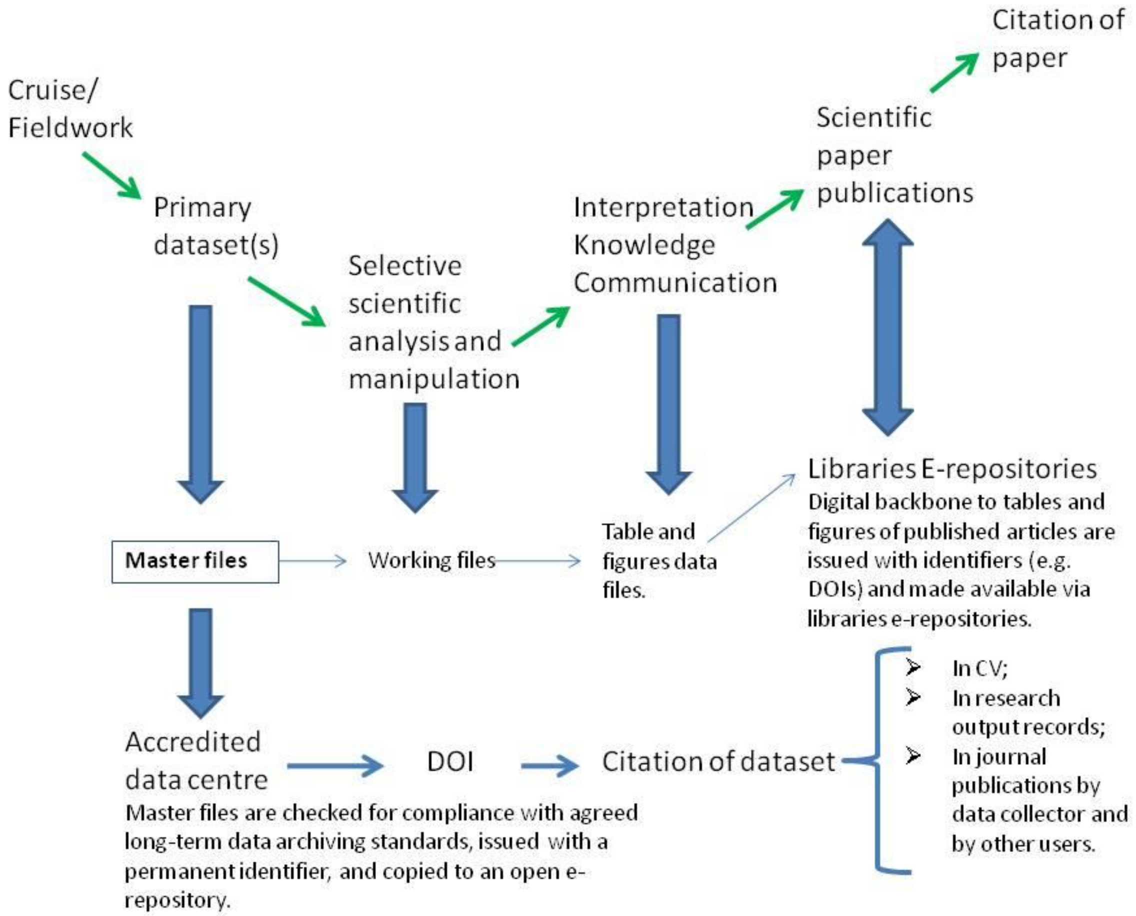
The datasets included in PDL will be linked from metadata records in the **Published Ocean Data repository (POD)** (current URL: http://193.190.8.15/pod/ ; future URL: http://www.publishedoceandata.org) providing a searchable catalogue of the available data sets.

Other similar models for data publication include Dryad, Pangaea and DSpace@MIT. Some are individual data repositories and others an aggregation of member nodes. The sharing of records through harvesting could also provide greater exposure for data exchange and the advent of data journals is an additional publication venue.

An interesting discussion has started in the informatics community – should we be using the metaphor "data publication" [51]. 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*." The conclusion is that no one metaphor suits all systems or methods. For the purposes of the Use Cases presented here, Data Publication is currently the term that best describes the process and is understood by the research community, but this may change in the future as the discussion continues.

## 7.2 From Fieldwork to Citation

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 4, key elements of the flow are indentified and should be key areas for collaboration and improvement in the coming 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
Master files
Working files
Table and figures data files.
Accredited data centre
DOI
Citation of dataset
In CV;
In research output records;
In journal publications by data collector and by other users.

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 are checked for compliance with agreed long-term data archiving standards, issued with a permanent identifier, and copied to an open e-repository.

Figure 4. End-to-end flow of data and information going from collection to publishing of data.

# 8.0 Recommendations

## 8.1 Interoperability/Standards Recommendations

Within domains, standards for 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 "universal" spoken language, commerce has been a major force in encouraging people to learn enough English to trade. The ubiquitous Microsoft and a few similar software giants have similarly forced compliance with their standard formats. Leaders who dominate the market establish an ad hoc standard. If Microsoft is added to Adobe, MathWorks, and half a dozen other software giants, 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, is the path that is presently recommended, much as in spoken language, Creoles develop and permit groups of mutually non-understanding people who must live and work together to communicate. Data "creoles" may bridge the interoperability until everyone adopts the commercial standard, be it ever so ad hoc.

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. Further development, implementation and uptake of brokering middleware is recommended as an important step forward. The Ocean research community, with its wide and multi-disciplinary diversity is an excellent test best for such implementation demonstrations.

The ACSI survey contained information about issues that straddle the boundary between data access and data quality:

- Users had problems finding data they knew were accessible;
- Users had problems finding documentation for datasets;

- Documentation quality was often poor;
- Network reliability was often poor and users would benefit from more robust protocols (of any sort);
- Data formats were not appropriate for more users; and
- Users were not aware of technical support services that were available.

Better understanding of these issues is important. A working group should be formed and should consider and recommend steps for both the technical and social/outreach aspects of the above issues. This includes outreach and education.

Outreach and Capacity Building for users to make them sophisticated enough to access data are needed, to enable them to seek support services. The ability of users to feel comfortable in a cross-domain environment is essential to further collaboration and address the complexities of global issues. Such activities should be built into the adoption and acceptance of Open Data.

## 8.2 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, even with volunteer reviewers, still requires substantial fiscal resources. Government support is pervasive throughout the research environment, covering infrastructure, salaries, university research activities, data management, publishing and community exchanges. A lot 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, publishing is moving journal support 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 7.1 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 disciplinary systems;
- Works across physical, social and economic sciences; and
- Promotes Access and use by the public and policy makers.

Metrics should be established to monitor progress in these areas. A Committee should be created with broad 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.

### 8.3 Data Publication/Data Citation Recommendations

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.

An accepted peer review methodology for datasets and/or data repositories has been discussed in the data management community at meetings like the 2012 Fall American Geophysical Union Meeting. This is an essential step.

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.

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.

Adoption of Digital Object Identifiers or equivalent “globally unique persistent identifiers” should be expanded and widely implemented.

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 OceanObsNetwork RCN could support such collaboration.

## Acronyms and abbreviations

| | |
|---|---|
| HTML | hypertext markup language |
| CSV | comma separated values |
| XML | extensible markup language |
| RDF | resource description framework |
| URIs | uniform resource identifier |
| HTTP | hypertext transfer protocol |
| XMPP | extensible messaging and presence protocol |
| DITA | Darwin information typing architecture |
| ASCII | American Standard Code for Information Interchange |
| ANSI | American National Standards Institute |
| UCS | universal character set |
| UTF-8 | UCS transformation format—8-bit |
| PNG | portable network graphics [16] |
| JPG | joint photographic experts group [16] |
| PDF | portable document format [16] |
| WSM | web map service [17] |
| WCS | web coverage service [17] |
| OGC | open geospatial consortium [17] |
| ISO | international organization for standards [18] |
| DAP | data access protocol [19] |
| WFS | web feature service [20] |
| WPS | web processing service [21] |
| GeoTiff | a public domain metadata standard which allows georeferencing information to be embedded within a TIFF file [22] |
| Shapefile | an ESRI proprietary yet popular geospatial vector data format for geographic information systems [23] |
| NetCDF | network common data form [24] |
| KMZ | keyhole markup language [25] |
| CEOS QA4EO | GEO/CEOS Workshop on Quality Assurance of Calibration & Validation Processes [27] |
| GEO | group on earth observations |
| CEOS | committee on earth observations satellites |
| GEOSS | global earth observation system of systems |
| QA4EO | quality assurance for earth observation [28] |
| GEOVIQUA | QUAlity aware VIsualisation for the Global Earth Observation system of systems [29] |
| QARTOD | quality assurance of real time oceanographic data [52 |
| Interoperability | the ability of diverse systems and organizations to work together [31] |

## References

[1] http://opendatahandbook.org/en/what-is-open-data/

[2] http://europa.eu/rapid/press-release_IP-11-1524_en.htm (Accessed 8 March 2013)

[3] http://www.nsf.gov/news/news_summ.jsp?cntn_id=123607 (Accessed 8 March 2013)

[4] http://www.nsf.gov/geo/earthcube/

[5] http://semanticweb.com/japan-embraces-open-data-launches-multiple-open-projects_b35158 (Accessed 8 March 2013)

[6] http://mashupaustralia.org/data-sources/ (Accessed 8 March 2013)

[7] http://onlinelibrary.wiley.com/doi/10.1002/asi.22634/abstract (Accessed 8 March 2013)

[8] http://wattsupwiththat.com/tag/climatic-research-unit-email-controversy/ (Accessed 8 March 2013)

[9] http://myip.ms/info/whois/207.97.227.245/k/3699568663/website/opendatahandbook.org (Accessed 8 March 2013)

[10]http://www2012.wwwconference.org/proceedings/nocompanion/wwwwebsci2012_braunschweig.pdf (Accessed 8 March 2013)

[11] http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/281 (Accessed 8 March 2013)

[12]https://www.google.it/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&ved=0CDEQFjAB&url=http%3A%2F%2Fciteseerx.ist.psu.edu%2Fviewdoc%2Fdownload%3Fdoi%3D10.1.1.71.8618%26rep%3Drep1%26type%3Dpdf&ei=zMs5UdqIEsmk4AT7iYD4Dw&usg=AFQjCNG_GStm5g2Z6Aig7bTIqsQKOkdbvQ&bvm=bv.43287494,d.bGE (Accessed 8 March 2013)

[13] http://www.ctg.albany.edu/publications/journals/hicss_2012_datasharing (Accessed 8 March 2013)

[14] http://www.surf.nl/en/projecten/pages/opendataandpublications.aspx (Accessed 8 March 2013)

[15] Nielsen, M. 2011 “Reinventing Discovery: The New Era of Networked Science” (2011), Princeton University Press, 280pp., ISBN: 9780691148908.

[16] Leadbeater, C., http://www.charlesleadbeater.net/cms/xstandard/Human_Genome.pdf

[17] http://opendefinition.org/okd/

[18] Palfrey, J. and U. Gasser, (2012) "Interop: The Promise and Perils of Highly Interconnected Systems," available: http://cyber.law.harvard.edu/publications/2012/interop (Accessed 8 March 2013)

[19] http://www.opendap.org/pdf/ESE-RFC-004v1.1.pdf (Accessed 11 Nov 2012)

[20] http://qa4eo.org/workshop_washington08.html (Accessed 11 Nov 2012)

[21] http://www.earthobservations.org/documents/committees/adc/200909_11thADC/DA-09-01a%20QA4EO.pdf (Accessed 12 Nov 2012)

[22] http://www.geoviqua.org/ (Accessed 12 Nov 2012)

[23] http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

[24] Whitfield, P. 2012, Canadian Water Resources Journal, vol 37 issue 1 pp. 23-36 DOI, 10.4296/cwrj3701866

[25] “NSB 11-79 Digital Research Data Sharing and Management” Report of the Task Force on Data Policies, National Science Board, December 2011

[26] Infrastructure for Spatial Information in the European Community (INSPIRE), http://inspire.jrc.ec.europa.eu/
[27] Global Monitoring for Environment and Security (GMES), http://www.gmes.info/
[28] National Spatial Data Infrastructure (NSDI), http://www.fgdc.gov/nsdi/nsdi.html
[29] NSF Data Observation Network for Earth (DataOne), https://www.dataone.org/about
[30] Jay Pearlman and Ryosuke Shibasaki,, "Global Earth Observation System of Systems", IEEE Systems Journal, Vol. 2, Number 3, September 2008; Global Earth Observation System of Systems (GEOSS), http://www.earthobservations.org/
[31] http://www.**wmo**.int/pages/prog/www/wis/
[32] http://www.gbif.org
[33] http://www.seadatanet.org
[34] http://his.cuahsi.org
[35] http://www.iode.org
[36] http://www.onegeology.org
[37]http://ec.europa.eu/research/infrastructures/pdf/workshop_october_2011/16_esfri_epos_cocco.pdf
[38] http://www.earthobservations.org/geobon.shtml
[39] European Commission, Communication from the Commission to the Council and the European Parliament: interoperability for Pan-European eGovernment Services. COM (2006), 45 final, Brussels, 2006.
[40] Turnitsa, C.D. (2005). Extending the Levels of Conceptual Interoperability Model. Proceedings IEEE Summer Computer Simulation Conference, IEEE CS Press]
[41] Nativi, S M. Craglia and J. Pearlman, 2013 "Earth Science Infrastructures Interoperability: the Brokering Approach" Journal of Selected Topics in Applied Earth Observation and Remote Sensing, in press.
[42] http://onlinelibrary.wiley.com/doi/10.1002/asi.22634/abstract
[43] http://www.dataone.org
[44] http://datafed.org
[45] http://www.opengeospatial.org
[46] http://ec.europa.eu/enterprise/newsroom/cf/_getdocument.cfm?doc_id=7140 (Accessed 8 March 2013)
[47] http://creativecommons.org/ (Accessed 8 March 2013)
[48] http://www.mendeley.com/catalog/citation-peer-review-data-moving-towards-formal-data-publication/ (Accessed 8 March 2013)
[49] JoRD: Journal Research Data Policy Bank Project **-** http://jordproject**.**wordpress**.**com/. (Accessed 24 Sep 2012)
[50] Tim Berners-Lee on "the next Web" at TED 2009 Conference. - http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html. (Accessed 11 Mar 2013)
[51] Parson, M. and P. Fox, **(**2013), "Is data publication the right metaphor". *Data Science Journal,* 10, 32-46. Available*:* http://dx.doi.org/10.2481/dsj.WDS-042 . (Accessed 21 Feb 2013)
[52] http://www.ioos.gov/qartod/welcome.html (Accessed 12 Nov 2012)

# General References

## Data and Business Models

**Bayh-Dole Act**. (Last modified 24 Jul 2012)
Available: http://en.wikipedia.org/wiki/Bayh–Dole_Act.

**Big data**: Crunching the numbers. *The Economist*, 18 May 2012.

**Borgman, Christine L**. (2012)
The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology* 63(6),1059-1078. doi:10.1002/ASI.22634.

**BSD Licenses**. (Last modified 27 Sep 2012)
Available: http://en.wikipedia.org/wiki/BSD_licenses.

**Climate Research Unit email Controversy**. (Last modified 1 Oct 2012)
Available : http://en.wikipedia.org/wiki/Climatic_Research_Unit_email_controversy.

**Creative Commons**. (Last modified 4 Oct 2012)
Available: http://creativecommons.org.

**Division of Ocean Sciences Data and Sample Policy.** (Last modified 2004)
Available: http://www.nsf.gov/pubs/2004/nsf04004/nsf04004.pdf.

**Free Software Foundation**. (Last modified 9 Oct 2012)
Available: http://www.fsf.org.

**Gleick, James** (2011)
*The Information*: *a history; a theory; a flood*. New York: Pantheon Books

**GNU General Public License (GPL)**. (Last modified 20 Jun 2012)
Available: http://www.gnu.org/licenses/gpl.html.

**INSPIRE**. (Last modified 15 Nov 2012)
Available: http://www.opengeospatial.org/pressroom/marketreport/inspire.

**Kossinets, Gueorgi, and Duncan J. Watts** (2009)
Origins of Homophily in an Evolving Social Network. *American Journal of Sociology*, 115, 405–50. doi:10.1086/599247. (Accessed 28 Feb 2010.)

**Mervis, Jeffrey** (2012)
Agencies Rally to Tackle Big Data. *Science,* 336, p.22.

**Open Geospatial Consortium**. (Last modified 15 Nov 2012)
Available: http://www.opengeospatial.org/.

**Nielsen, Michael** (2012)
*Reinventing Discovery: The New Era of Networked Science*. Princeton and Oxford: Princeton University Press.

**Palfrey, John, and Urs Gasser** (2012)
*Interop: The Promise and Perils of Highly Interconnected Systems.* New York: Basic Books/Perseus.

**Policy and Award Policies and Procedures Guide**. (Last modified 18 Jan 2012)
Available: http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/nsf11_1.pdf.

**Summary White Paper**. (Last modified 16-17 Nov 2011)
ftp://ftp.earthobservations.org/ExCom/22/08_Report%20of%20the%20Data%20Sharing%20Task%20Force.pdf.

**What is the OGC?** (Last modified 30 Jan 2012)
Available: http://youtube.com/.

**Whitman, Walt** (1919)
*What whispers are these; Leaves of Grass.* Garden City, N.Y: Doubleday.

## Data Publication/Data Citation

**Ball, Alex and Duke, Monika** (2012)
*How to cite data sets and link to publications*. Edinburgh, UK: Digital Curation Centre. (DCC How-to Guides). Available: http://www.dcc.ac.uk/webfm_send/525. (Accessed 11 Sep 2012)

**Beagrie, Neil, Beagrie, Robert and Rowlands, Ian** (2009)
Research Data Preservation and Access: The Views of Researchers. *Ariadne*, Issue 60, 30 July 2009. Available: http://www.ariadne.ac.uk/issue60/beagrie-et-al. (Accessed 11 Sep 2012)

**Blue Ribbon Task Force on Sustainable Digital Preservation and Access** (2010)
*Sustainable economics for a digital project: ensuring long-term access to digital information. Final report edited by Amy Smith Rumsey.* 110pp. Available: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf. (Accessed 11 Sep 2012)

**Callaghan, Sarah, Lowry, Roy and Walton, David** (2012)
Data Citation and Publication by NERC's Environmental Data Centres. *Ariadne,* Issue 68 .
Available: http://www.ariadne.ac.uk/issue68/callaghan-et-al. (Accessed 14 Sep 2012)

**CODATA Data Science Journal**
Available: http://www.codata.org/dsj/. (Accessed 14 Sep 2012)

***Data's shameful neglect: editorial***. *Nature*, 461(145): 168-170. (2009).
Available: http://www.nature.com/nature/journal/v461/n7261/full/461145a.html. (Accessed 14 Sep 2012)

**Digital Object Identifier**
Available: http://www.doi.org/. (Accessed 14 Sep 2012)

**Dryad**
Available: http://datadryad.org/. (Accessed 14 Sep 2012)

**Earth System Science Data: the data publishing journal**.
Available: http://earth-system-science-data.net/. (Accessed 14 Sep 2012)

**GeoScience Data Journal**
Available: http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%292049-6060. (Accessed 20 Feb 2013)

**Harper, Corey A. (Ed)** (2012)
Linked data for Libraries, Archives and Museums. *Information Standards Quarterly*, 24(2/3).
Available: http://www.niso.org/apps/group_public/download.php/9422/isqv24no2-3.pdf. (Accessed 12 Sep 2012)

**High Level Expert Group on Scientific Data [Wood, J. et al]** (2010)
*Riding the wave: how Europe can gain from the rising tide of scientific data*. Final Report. 36pp. Available: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf. (Accessed 11 Sep 2012)

**JoRD: Journal Research Data Policy Bank Project**
Available: http://jordproject.wordpress.com/. (Accessed 24 Sep 2012)

**Lawrence, B., Jones, C., Matthews, B., Pepler, S. and Callaghan, S.** (2011)
Citation and Peer Review of Data: moving towards formal data publication. *International Journal of Digital Curation*, 6(2), 4-37. doi:10.2218/ijdc.v6i2.205. Available: http://www.ijdc.net/index.php/ijdc/article/view/181/265 . (Accessed 14 Sep 2012)

**National Science Foundation Data Management Plan**
Available: http://www.nsf.gov/bfa/dias/policy/dmp.jsp. (Accessed 14 Sep 2012)

**Pangaea**
Available: http://www.pangaea.de/. (Accessed 11 Sep 2012)

**Parson, M. and Fox, P. (**2013)
Is data publication the right metaphor. *Data Science Journal,* 10, 32-46. Available*:* http://dx.doi.org/10.2481/dsj.WDS-042 . (Accessed 21 Feb 2013)

**Penev, L., Mietchen, D., Chavan, V., Hagedorn, G., Remsen, D., Smith, V and Shotton, D.** (2011)
*Policies and Guidelines for Biodiversity Data*. Pensoft Publishers, 34pp. Available: http://www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf. (Accessed 14 Sep 2012)

**Piwowar, Heather A., Vision, Todd J. and Whitlock, Michael C.** (2011)
Data archiving is a good investment. *Nature*, 473, p.285 . doi:10.1038/473285a. Available: http://www.nature.com/nature/journal/v473/n7347/full/473285a.html. (Accessed 11 Sep 2012)

**Piwowar, H.A., Day, R.S. and Fridsma, D.B** . (2007)
Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE,* 2(3): e308. doi:10.1371/journal.pone.0000308. Available: http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0000308. (Accessed 14 Sep 2012)

**Published Data Library (PDL)**
Available**:** https://www.bodc.ac.uk/data/published_data_library/ (Accessed 11 Sep 2012)

**Ruuselepp, Ruivo** (2008)
*Infrastructure planning and data curation : a comparative study of international approaches to enabling the sharing of research data.* Edinburgh, UK: Digital Curation Centre & JISC, 108pp. Available: http://www.jisc.ac.uk/media/documents/programmes/preservation/national_data_sharing_report_final.pdf . (Accessed 11 Sep 2012)

**Swan, Alma and Brown, Sheridan (**2008)
*To Share or not to share* **:** *publication and quality assurance of research data outputs.* London*,* UK*:* Research Information Network*,* 56pp. Available: http://rinarchive.jisc-collections.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs (Accessed 28 Sep 2012)