Article

JIS

Journal of Information Science
37(6) 555–569

DOI: 10.1177/0165551511412705
jis.sagepub.com
SAGE

A conceptual framework for managing very diverse data for complex, interdisciplinary science

Mark A. Parsons
National Snow and Ice Data Center, University of Colorado, USA

Øystein Godøy
Norwegian Meteorological Institute, Norway

Ellsworth LeDrew
University of Waterloo, Canada

Taco F. de Bruin
NIOZ Royal Netherlands Institute for Sea Research, The Netherlands

Bruno Danis
Antarctic Biodiversity Information Facility, Belgium

Scott Tomlinson
Indian and Northern Affairs Canada, Canada

David Carlson
UNAVCO, USA

Abstract

Much attention has been given to the challenges of handling massive data volumes in modern data-intensive science. This paper examines an equally daunting challenge – the diversity of interdisciplinary data, notably research data, and the need to interrelate these data to understand complex systemic problems such as environmental change and its impact. We use the experience of the International Polar Year 2007–8 (IPY) as a case study to examine data management approaches seeking to address issues around complex interdisciplinary science. We find that, while technology is a critical factor in addressing the interdisciplinary dimension of the data intensive science, the technologies developing for exa-scale data volumes differ from those that are needed for extremely distributed and heterogeneous data. Research data will continue to be highly heterogeneous and distributed and will require technologies to be much simpler and more flexible. More importantly, there is a need for both technical and cultural adaptation. We describe a vision of discoverable, open, linked, useful, and safe collections of data, organized and curated using the best principles and practices of information and library science. This vision provides a framework for our discussion and leads us to suggest several short- and long-term strategies to facilitate a socio-technical evolution in the overall science data ecosystem.

Corresponding author:
Mark A. Parsons, UCB449, University of Colorado, Boulder, CO 80309, USA.
Email: parsonsm@nsidc.org.

Keywords
Antarctic; Arctic; data archives; data casting; data discovery; data documentation; data management; data preservation; data sharing; interdisciplinary science; international collaboration; International Polar Year; linked data; long tail of science; metadata; ontology; open data; polar regions; research data; socio-technical systems; standards; World Data System

1. Introduction

That modern scientists face a 'data deluge' has almost become a cliché. Jim Gray and other top computer scientists have described how the techniques and technologies needed to perform data-intensive science may now comprise a new 'fourth paradigm' of research [1]. Scientists have long recognized experimental and theoretical science as the two basic research paradigms. In the latter half of the twentieth century, computer simulations emerged as a third paradigm for scientists to explore domains that are inaccessible to theory and experiment such as predicting global climate change. These simulations and attendant enhanced measurement techniques produce ever more data and lead us to Gray's fourth paradigm. This is a valuable insight; but to date, the e-science or data-intensive science emphasis has been on dealing with overwhelming data volumes [2–5]. For example, the UK Research Council defines e-science as 'large-scale science carried out through distributed global collaborations enabled by networks, requiring access to *very large data collections, very large-scale computing resources*, and high-performance visualization' [5 p. 69, our emphasis]. Similarly, initial investments in cyberinfrastructure in the USA, such as the TeraGrid,[1] were geared toward large-scale data systems and computation. This emphasis on huge data volumes has underplayed another dimension of the fourth paradigm that presents an equally daunting challenge – the *diversity* of interdisciplinary data and the need to interrelate these data to understand complex systemic problems such as environmental change and its impact.

The International Polar Year (IPY) was a US$1.2 billion investment in polar research involving 50,000 participants from 60 nations [6]. It emphasized both interdisciplinary collaboration and broad understanding of complex, intertwined, physical, biological, health, and social systems. In IPY, scientists collected every possible form and format of data: images wide, narrow and panoramic, profiles upward and downward, hourly to millennial time series, isotope ratios and fractions, energy and material fluxes, species identification and distributions, interviews in common and rare languages, disease types and rates, genealogies and genetic sequences, samples and artefacts, singular events and gradual processes, and so on. (For a more complete description of IPY and its data see [7].) IPY also coincided with the Electronic Geophysical Year, which sought to promote open and excellent data stewardship of Earth science data [8]. IPY and its data provide an excellent case study to understand the interdisciplinary dimension of the fourth paradigm. In this paper, we discuss the experience of data scientists seeking to collect, curate, and provide data from IPY and related programmes to address interdisciplinary science goals.

A major challenge of interdisciplinary, systemic research is in the use of so-called research collections. The National Science Board (NSB) defines three broad categories of data collection: research, resource, and reference collections [9]. Research collections, which are data collected by individual investigators and small research groups, stand out as critically absent from current data systems. These unique observations of the Earth system are unrepeatable and increase in value over time, but they are underutilized, vanishing, and often forgotten. Research collections are especially critical in polar research where making routine observations is challenging because of frequently severe weather, limited access and communications, extended darkness, extreme stresses on instrumentation and data storage, and unique orbital and cloud limitations to remote sensing. Research collections present some special challenges:

- They lack established or standardized data systems.
- They are not broadly shared or discoverable and are, therefore, little used beyond their original application.
- They are often project specific and thus not well integrated or usable in conjunction with more standardized resource and reference collections.
- They are growing in size and complexity, as researchers develop and adopt new technologies such as unpiloted aerial vehicles, telemetric sensor networks, sensor webs with adaptive sampling systems, and even nanotechnologies.
- Unlike large remote-sensing programmes, which are usually packaged with strong data distribution programmes, the data management needs of research collections are not well funded or planned.
- They are at the greatest risk of loss [9].

Heidorn [10] describes these data as the long tail of science – the small, but very diverse data sets collected by the majority of scientists. It is this heterogeneity of data and the cultural diversity of their collectors, not their volume, that present the greatest challenges. Furthermore, these research data need to be integrated with reference and community

data, and that requires adaptation by both data managers and collectors. As data managers for IPY, we find that, while technology is a critical factor to addressing the interdisciplinary dimension of the fourth paradigm, the technologies developing for exa-scale data volumes differ from those needed for extremely distributed and heterogeneous data. Furthermore, as with any socio-technical change, the greater challenges are more socio-cultural than technical [cf. 11].

We have established a working vision to guide our efforts that we believe could be a useful vision for science in general. Simply put, we believe that data should be discoverable, open, linked, useful, and safe. More specifically, we believe that data should be:

- *Discoverable*. Data must be capable of being located, identified, and generally assessed through simple tools available to many communities.
- *Open*. Data should generally be openly accessible. We find the current Wikipedia definition describes our intentions well: 'Open data is a philosophy and practice requiring that certain data be freely available to everyone, without restrictions from copyright, patents or other mechanisms of control.'
- *Linked*. Data should be interrelated and connected. There are many ways to interconnect data and we use the term 'linked data' rather generally. Increasingly the term refers to connecting data via dereferenceable URIs on the web.
- *Useful*. Data must be able to be used for a practical, advantageous purpose or in several ways by defined but possibly very different users.
- *Safe.* Data should be protected from risk, corruption, and loss; now and over the long term.

These are simple terms or concepts, but they sum up the primary objectives we seek in managing diverse scientific data. They are in keeping with the trends and principles outlined by the Association of Research Libraries' *Agenda for Developing E-Science in Research Libraries* [2] and other community guidance [12, 13]. We do not offer new principles, per se, but rather a different perspective that is helpful in understanding the complexities of extremely diverse data. By reducing data stewardship to simple terms, we can more easily see the full context of the enterprise. While this oversimplification certainly misses some detailed aspects of data stewardship, the model affords a more holistic view, which we have found useful in understanding interdisciplinary issues. Beers and Bots [14] assert that promoting interdisciplinarity requires institutional changes and/or a lowering of transaction costs. While they take an information modelling approach to address the transaction cost issue, we focus more broadly.

Our perspective is similar to Star and Griesemer's 'ecological analysis'.

> An advantage of the ecological analysis is that it does not presuppose an epistemological primacy for any one viewpoint; the viewpoint of the amateurs is not inherently better or worse than that of the professionals, for instance. We are persuaded by Latour that the important questions concern the flow of objects and concepts through the network of participating allies and social worlds. [15, p. 389]

We consider what Star and Ruhleder call an 'ecology of infrastructure', where the term ecology 'refers to the delicate balance of language and practice across communities and parts of organizations; it draws attention to that balance (or lack of it)' [16, p. 117]. We examine what we call the 'data ecosystem' – the people and technologies collecting, handling, and using the data and the *interactions* between them. Data managers emphasize the need to consider the entire data lifecycle [e.g. 2, 17] and we consider the data ecosystem a natural extension of that concept. It is similar to what Nardi and O'Day call 'information ecologies' [18]. Yet, whereas they emphasize the importance of locality and focus on micro-scale interactions between people and technology, we take a broader, more systemic view. Their examples of information ecologies are local units such as an individual library or a hospital intensive care unit. We explore not only interactions within an ecology, but also how these ecologies interact across broader, even global, scales. Nardi and O'Day note how 'the ecology metaphor provides a distinctive, powerful set of organizing properties around which to have conversations' [18, p. 50]. We agree and further extend the metaphor and conversation to the whole ecosystem.

The ecosystem concept is a useful metaphor because ecosystems are adaptive in complex Darwinian ways. They have myriad niches that emerge, co-evolve, and survive or senesce depending on environmental context. Ecosystems are considered most healthy when diverse. They take time to grow, and they evolve in response to interaction with all their components. Crucially, ecosystems, like infrastructure [16], comprise relationships and interconnections between and amongst people, technologies, and data. It is a complex system involving data collectors, stewards, and users as well as sponsors and stakeholders; emergent and historical transparent technologies; and ever-growing data along with their myriad associated artefacts. The system must be understood in totality in order to optimize the whole and not just the individual components.

We note, however, that we are practitioners, not theoreticians. We are interested in how this holistic, ecological view can inform our day-to-day lives as actors in the data ecosystem stewarding and communicating about very diverse data. We participate in the system but also provide a driving force with conscious intention to design a system to meet particular needs and objectives. We, therefore, use our basic model of discoverable, open, linked, useful and safe as an outline to discuss specific observations while maintaining a big picture view of the sociotechnical culture, systems, and strategies necessary to address the heterogeneous, interdisciplinary dimension of the fourth paradigm. We conclude by suggesting several short- and long-term strategies to facilitate a socio-technical evolution in the overall science data ecosystem.

2. Data should be: discoverable

For data to be used beyond its original purpose, it must be discoverable within and across disciplines. In our experience, most researchers find data primarily through personal connections or through discipline-specific, community data systems. To date, most approaches to facilitate data discovery require the creation of formally structured descriptions, i.e. metadata records, that are then submitted to centralized registries like the Group on Earth Observations (GEO) data registry, WMO Information System (WIS), Global Biodiversity Information Facility (GBIF), Global Change Master Directory (GCMD), or the forthcoming European INSPIRE infrastructure for data. Many of these initiatives are driven at high intergovernmental levels (e.g. WIS, INSPIRE, and GEO) and sometimes lack a clearly designated scientific user community or community of practice to engage in and provide feedback on the development of systems and tools and their implementation. Often a goal of this registry-based approach is to create a single portal or 'one-stop shop' that describes all data. While some IPY participants found value in creating a unified portal at the GCMD, it is clear that this approach is lacking in several ways and will not adequately serve all disciplines. For example, it requires all disciplines to adhere to a common metadata format and, more critically, a particular controlled vocabulary for describing the content of the data. This approach does not recognize that different disciplines have very different ways of conceiving of intellectual concepts that are reflected in how they structure, represent, and describe their data. For example, a climate researcher, an operational shipping forecaster, an Arctic wildlife ecologist, and a local hunter all have a keen interest in determining the edge of the sea ice, yet each defines the 'edge' and describes its characteristics in a different way and in a different data structure. Furthermore, there is even debate within disciplines about how to represent and describe concepts. This is particularly evident in biological taxonomies, which are constantly under revision with great debate about the detailed specifics. In short, registries work well for discovery when they are describing relatively similar objects, but they become unwieldy and imprecise when describing heterogeneous objects to diverse specialized audiences.

Rather than a one-stop shop, a better metaphor is a marketplace or bazaar – a virtual space where all data can be found, but specialist portals provide the expertise, information, and referrals necessary to identify and understand data within a specific disciplinary context. Returning to the ecosystem metaphor, each of these specialists can be seen as filling different niches by providing and exchanging specialized services. At a practical level, this implies that disciplines and research or decision focus areas need to develop portals to meet their needs. It also means that data centres need to expose their data and metadata through multiple protocols that allow these different systems to automatically identify and acquire the information to be presented in ways that are meaningful to their designated user communities. IPY has had initial success creating a union catalogue where multiple data centres expose their metadata through the Open Archives Initiative Protocol for Metadata Harvesting. The GCMD serves as an overall authority catalogue in this initial system and harvests all available metadata, but, more importantly, individual archives can choose to harvest and expose only select data in ways relevant to their users. Evolving examples of these specialized portals range from a global cryospheric portal to a portal offering a local and traditional knowledge perspective of biodiversity change.

A new initiative, the Polar Information Commons,[2] takes this approach a step further. It explores an approach where a small, machine-readable 'badge' is attached to the metadata or data. This badge asserts that the data are open and allows generic search engines or customized portals to automatically identify and locate relevant data. This can be coupled with a data-advertising approach called 'data casting'. Data casts extend the popular RSS and ATOM subscription services to include geospatial, temporal, and other attributes commonly used in the sciences. Data providers can then advertise their data through a web feed, and users can then subscribe to data feeds and view them through any feed reader. Better yet, tailored aggregation services can be developed that restrict results to certain types of data, spatial extents or temporal ranges, etc. Data casting can be implemented at a data set and at an item level and can provide the ability for the actual data to be directly retrieved as a result of subscribing to the data cast. Data casting is increasingly used in Earth science. The GEO Portal[3] aggregates geoRSS feeds and NASA has been supporting further development of data casting for both whole collections and individual files or records.

Another emerging technology that facilitates the marketplace style of data discovery is OpenSearch.[4] OpenSearch is a set of simple formats used for sharing search results in a federated system. OpenSearch was developed by a subsidiary of

Amazon and has been very successful in the commercial world (it underpins the Amazon Marketplace), but is only now beginning to be used in science. An advantage of OpenSearch is that it provides a way for data stewards to present information that would be otherwise inaccessible to search engines and web crawlers. More importantly, it allows those most familiar with the content in question to enable search in a way that is most appropriate. Directly involving this subject knowledge view is likely to increase the relevance of search results [19]. All told, data casting and OpenSearch allow for greater distribution of data sources and more tailored approaches to presenting the data – two attributes especially helpful to discovery of diverse, complex data.

Of course, for any discovery system to work well, it is necessary to obtain richer, more detailed metadata, generated in many ways, so that discovery systems can identify and access particular measurements within broader collections. Given the lack of routine, automated metadata collection with research collections, data managers and computer scientists need to use creative means to automatically capture metadata from other sources – proposals, abstracts, publications, conferences, etc. – as well as directly from the data themselves. That said, the knowledge of the original data creator is paramount. Regardless of the technical approach, there must be education of the scientists in the peculiarities of metadata. This is particularly critical with research data collections, which so often lack formal data systems and rely on the scientists themselves to describe and make the data available. Early in the IPY, at data management meetings held to enable effective discovery, we found that metadata was considered a fearsome topic that would drain the time of the scientists and was thus to be avoided. Simple tools such as web-based templates with drop-down lists and on-line tutorials, along with substantial personal assistance helped ease the process. As automation increases, the job can be made easier but willingness and ease for the data collector to express and share knowledge about data particularities will always be essential. One positive outcome of this experience garnered through IPY and elsewhere has been support for development of curricula in science data management. This may be provided through block courses, on-line tutorials, or as a more formal component of graduate degree programmes. Several promising initiatives are underway and being viewed with interest.

3. Data should be: open

Closely related to data discovery is open access. The IPY data policy emphasized open and timely access to most data. This principle builds from the emerging consensus on the value of open data and the general movement toward more openness and less restriction, especially in the physical and life sciences [20–26]. Some argue that freedom of access to information is fundamental to participation in a knowledge society [27] and that open access should be the default rule not the exception [12]. International bodies such as the World Meteorological Organization, the International Council for Science and the Group on Earth Observations have long advocated for 'full and open access' to research data [28–30]. Implementation of this principle of full and open access is much more variable at the national level, though, and in some nations the national policy is entirely inconsistent with the principle. Again, research data collections are often the most likely to be unavailable. A recent PARSE.Insight (Insight into issues of Permanent Access to the Records of Science in Europe) survey suggests that only 25 per cent of research data across many disciplines are openly available [31]. Furthermore, there is huge variability in attitudes towards data sharing across research disciplines [32]. As an example, in the field of sociology, Nicholson and Bennett note: 'Our observations indicate that new researchers neither provide nor have access to the full range of research data collections, thus impeding data reuse and replication opportunities for the advancement of knowledge' [33, p. 514].

Some of this restriction may be a result of past practice that encouraged embargoing of data until formal publication. Some of it may be related to perceived commercial value of data and the role of lawyers promoting agreements for intellectual property rights as a prudent safeguard. One of the attractive tenets of IPY, however, was that rapid sharing of data would encourage new insights as a result of the intrinsic interdisciplinarity of the processes in the region. Open access to data in the shortest feasible time scale was seen as necessary to meet IPY objectives, especially in a time of rapid change. We see this principle as fundamental to modern interdisciplinary science and as a catalyst that may speed the transformation to a more data-driven approach to science as well [cf. 11].

Open data also helps ensure accountability. The recent 'Climategate' controversy, in which senior climatologists were accused of manipulating important global temperature data, starkly illustrated the importance of open research data for the integrity of science. While multiple high level investigations cleared the researchers of any wilful wrongdoing, the investigations also emphasized the need for data to be more openly available to ensure credibility and avoid future misguided controversy [34–36]. Indeed, data transparency has proven an antidote to those rare instances of scientific fraud [37].

Reasons researchers give for denying data access include concern over data misuse, ethical or legal breaches (i.e. a lack of trust), and simply the effort necessary to make data available. In some disciplines, patents and other restrictions

are used as a purported way to provide financial incentive and support future research, but these incentives often do not live up to their financial promise, especially in academic research where restriction imposes more cost than reward [26]. More generally, we have found that a major challenge to increasing data sharing is rooted in the culture of science and the research academy. A researcher's merit is judged largely on the number and quality of their peer-reviewed publications. This one-dimensional metric provides little incentive to compile and document data beyond the needs of the original research. Indeed, it creates an incentive to restrict data access in order to maximize the number of publications a researcher can produce from the data. There is a cost to a researcher to sharing data, but others receive most of the benefits. Data creators can be leery of sharing data with 'outsiders'. This results in project or disciplinary data silos that hinder interdisciplinary research. Science and funding agencies should recognize the intellectual effort necessary to compile and document a good data set and provide incentives to produce and share good data. While we strongly encourage soft incentives like encouraging formal data citation as a means of crediting data authors [38], IPY found the strongest incentive is when data deposit in an open archive is a requirement for ongoing research funding, and that requirement is supported by identified, funded archives. Further, this type of strict enforcement is more effective when there is a good working relationship between the archive and the data provider.

Overall, we best serve science and society when we view data as a common, public good. This is the basic principle of an information commons. Creative Commons has proposed a *Protocol for Implementing Open Access Data*[5] that proposes that data be placed as fully as possible in the public domain while also asserting expected norms (not legal requirements) of ethical behaviour for users and providers of the data. An example of this approach is the developing Polar Information Commons, noted previously. The approach of an information commons is likely to be the simplest, most useful, and sustainable approach to addressing issues of open data access. On the other hand, IPY was careful to recognize legitimate restrictions to data access that protect privacy of human subjects, respect the intellectual property rights of local and traditional knowledge holders, and avoid situations where data release could cause harm. While we are strong advocates of open data, we recognize there are great complexities in the details. An evolving legal and social science research agenda is needed to best balance between society's need for open data and the need to protect people, heritage, endangered species, and cultures from misuse. Information science has begun to explore some of these issues in medicine [39], human rights [40], and locational privacy [41], but broader research more focused on data, rather than publications, is needed.

4. Data should be: linked

Once data are discoverable and open, they need to be readily associated with related and supporting data to enable insightful understanding and interdisciplinary research. This concept of sharing and connecting diverse pieces of data and information is often called 'linked data'. Currently, most Earth system science data are managed in hierarchical file systems and relational databases. Connections to or relationships with other data are through structured descriptions of hierarchies in metadata, XML schemas, and file structures or through primary key relationships defined in schemas for relational databases or networks of databases. An increasingly popular and more specific concept of linked data is through the semantic web. This linked data approach interrelates diverse data on the web through a 'graph database' model that does not rely on a fixed schema. Additional nodes or elements can easily be added (horizontally scalable). The Resource Description Framework (RDF) is a basic W3C standard that is used to describe the relationships in the graph database. RDF and other standards then facilitate the interconnection of data without relying on common schemas and hence allow schemas to evolve without requiring changes in data access systems.[6] The W3C Semantic Web Education and Outreach Interest Group has begun a 'Linking Open Data' community project to 'extend the Web with a data commons by publishing various open data sets as RDF on the Web and by setting RDF links between data items from different data sources'.[7] The inherent extensibility of the Linked Data approach would seem to serve complex, diverse data well, but Earth system science data are just beginning to appear in the Linked Open Data cloud.

As discussed above, polar science (indeed most science) has relied almost exclusively on metadata catalogues not only for data discovery but also to associate related data through hierarchies, structured relationships, and defined keywords or vocabularies. Unfortunately, while these registries describe the data well, they do not provide consistent access to the data. There may be a direct link to the data, but too often it is just a link to another web site or there is no link at all. So while metadata is increasingly linked across catalogs, the actual data remain extremely disconnected. Furthermore, the metadata catalogues are linked or federated primarily by agreeing on a common structure or syntax of the shared metadata. The linkage does not often capture or translate the meaning, or the semantics, of the actual content of the metadata. In IPY, we were able to agree very quickly on a basic, flexible metadata scheme to which existing systems could easily adapt and new systems could easily adopt. It was much more difficult to agree on standards for describing the data – the keywords, the vocabularies, the units, the 'semantics'.

Once again, this is most challenging with the data sets in the long tail of science. Large data systems with their high costs and large data volumes demand syntactical and some semantic consistency for the systems to function and for funding agencies to get suitable return on their large investment. These data systems handle relatively homogenous data and often have coherently constructed infrastructures (see for example, the World Climate Research Program Climate Model Intercomparison 3 Multi-Model Data Set[8]). They can, therefore, provide greater interoperability, at least within the domain in question. The highly disparate data in the long tail lack similar syntactical standards (e.g. common metadata formats) and their disciplinary diversity challenges consistent semantics (e.g. standard keywords). Conventional hierarchical taxonomies and keyword lists lack the detail and consistency to fully describe the range of data and how they are interrelated. Much richer semantics are necessary, ideally in the form of interconnected ontologies – formal, computer-readable definitions of concepts and their relationships. Ontologies enable the flexible networks of truly linked data to fully function [42].

This semantically rich, easily extensible, linked-data approach seems most suitable for interconnecting diverse research data across existing disciplinary data silos. It allows adaptive and iterative development, without the construction of complex hierarchical structures or heavyweight standards at the outset as is typically done in large-scale infrastructure projects like INSPIRE or GEOSS. However, linked data is an evolving approach that still requires substantial agreement on detailed standards, such as ontologies, to realize its full potential. As we will discuss below, standards development is largely a technical exercise, but the adoption of standards is a social exercise. Our experience suggests that, when developing data systems, it is best to start simple, using proven and known approaches, and then take an incremental, adaptive approach to expanding their interconnection. These interconnections can occur initially through geospatial or hierarchical approaches while the community may still explore more sophisticated semantic connections. Regardless of how data are interconnected, they still need additional context, quality assessment, and explanation to actually be useful beyond their original purpose.

5. Data should be: useful

Consider the variety of applications of IPY data by researchers, operators, and decision makers. Carbon impacts researchers, as an example, need information on northern terrestrial carbon sources and sinks; northern and southern oceanic carbon sinks and sources; and changes in hydrology and snow cover and their influence on vegetation and fire. Marine operators for commercial shipping and resource extraction, as another example, need predictions of ice, ocean, and marine weather in daily, monthly, seasonal and multi-year forecasts. Decision makers could include consumers and managers of polar marine and terrestrial ecosystem resources, health specialists, sociologists and community activists, economists, and local, regional and national governments. These users need information on marine and terrestrial ecosystems; community adaptability and survivability; polar biodiversity; health impacts of regional or global pollutants deposited in polar regions; condition, timing, and predictability of snow cover, lake and river ice, and sea ice; structural changes in permafrost; options and mechanisms for local management and governance.

When considering such a diversity of users, needs, and applications, it is clear that much more than just the data are necessary. Users need the data to be coherent in form and semantics with their models and analysis tools. They need rich documentation fully describing data uncertainties and fitness for use. They need context and background about algorithms, calibrations, and methods [43, 44]. Furthermore, the decision makers do not really need data at all but rather information presented in compelling, readily interpretable ways. Effective data and information display can encourage greater data sharing, increase understanding of complex processes, and enable wiser decisions [45–48].

Strictly speaking, the data need to be both ‘useful’ by meeting a need and ‘usable’ by being actually suitable for use. Davis, in describing human–computer interaction, makes the distinction between ‘perceived usefulness’ and ‘perceived ease of use’ [49]. In this interdisciplinary context we take a broad view and consider both concepts under our general term ‘useful’. Much of our discussion focuses on the usability or ease of use of data, but the greater challenges revolve around the usefulness of the data. There are great challenges to sharing sufficient context and provenance information across disciplinary boundaries for users to effectively evaluate and correctly apply data.

All this suggests, as has been stated elsewhere [2, 9, 29, 50], that there will be an increasing need for informatics specialists in the twentieth century. These specialists will need to include not only computer scientists and systems engineers grappling with complex technical issues, but also data scientists, data curators, librarians, data ‘wranglers’ [51], information designers, and even artists grappling with social systems and improving human understanding. Society will increasingly rely on those professionals who act as translators to make complex, distributed data accessible and useful. Training, development, and reward structures for these new professionals need to be top priorities. At the same time, scientists themselves need education in the fundamentals of data management through structured curricula. There has been less emphasis in this area, but scientists and data managers need to engage in the kind of learning that occurs

through ongoing dialogue and experimentation in order to develop the literacy necessary to address complex sociotechnical issues [16]. Close collaboration between data scientists and domain scientists is necessary to make a functional cyberinfrastructure. We found repeatedly in IPY that taking a simple yet adaptive approach to system development led to more efficient and more useful systems. A large EU-funded IPY project called DAMOCLES provides a good example of how simplicity and collaboration can lead to increased use of standards and greater efficiency while encouraging the collaborative investment necessary to develop interdisciplinary, socio-technical systems.

DAMOCLES (Developing Arctic Modelling and Observing Capabilities for Long-term Environmental Studies) included oceanography, meteorology, and other geosciences. The project involved both scientists and industry partners. There was a strong imperative for the data to be well documented and consistent across the project. Industry partners, in particular, were very interested in community standards for file formats. Ideally, documentation should fully describe who collected what data where and when and also include identification of the variables, their units, and their uncertainty. This is difficult in an interdisciplinary framework, and it is not easy within disciplines either. It is one thing to define standards. It is quite another to encourage actual use of the standards. Often different scientists have their own file format even when they are using the same instrument from the same provider. The scientist knows the contents of the files generated. Other users often do not. How can these two worlds communicate, but through standards?

Based on the experience of the oceanographic community, DAMOCLES investigators decided to standardize around NetCDF – a standard, hierarchical, self-describing format – because it is supported by many common analysis tools (e.g. Matlab, R, Ferret) and data access protocols like OpeNDAP that facilitate sub-setting before downloading and other features. Furthermore, scientists using NetCDF have been very active in defining the specific conventions of how to use the self-describing aspects of the format.[9] Consistent adherence to these conventions like the Climate and Forecast (CF) extensions make the data more useful for more people. CF extensions, however, focus on usage metadata describing variables, units, valid ranges etc. They do not cover discovery and contextual metadata describing who measured what, where, and when. In DAMOCLES, additional global attributes were added to NetCDF files compliant with the CF structure that later proved sufficient to fulfil the minimal discovery metadata requirements outlined by IPY.[10]

To encourage scientists to actually use the standards, the DAMOCLES data submission system included a compliance checker. It soon became apparent that the checker needed to be tuned to be less restrictive to help scientists adapt to the newly required format. This was especially true with communities less experienced with NetCDF. In a related project, biological scientists worked with data in Microsoft Excel, but during the conversion to NetCDF it became evident that they needed a more consistent structure for their spreadsheets. After the new structure was implemented in the Excel documents, data was easily transferred from the spreadsheet to NetCDF for inclusion in the national IPY archive in Norway. Future projects will also use the spreadsheet structure as a way to facilitate data and metadata transfer to archives.

Much of the incentive for standardizing the DAMOCLES file format was the desire for a low-cost data management system. Scientists often perceive data management as a funding competitor that draws money away from research, so there was an imperative to keep the data management costs low. To meet this imperative while still ensuring adequate data documentation and standardization, DAMOCLES scientists had to do more than they were accustomed to document and consistently format their data. It took time to build collaborative relationships between the data managers and data providers, but this ultimately turned out to be an efficient approach. We saw similar successes built around collaboration in other projects addressing ice cores, Arctic biodiversity, atmospheric science, indigenous knowledge, and other topics. It is an ongoing challenge to maintain that collaboration and to provide an easy data submission framework that also maintains data integrity and usability.

6. Data should be: safe

Keeping data safe, now and for the long term, is perhaps the best-understood but most difficult challenge of the fourth paradigm. By safe, we simply mean that data integrity is recorded and preserved and the data remain usable for future generations. They are safe from technical obsolescence and deterioration, safe from hacks or general undocumented change, safe from the loss of contextual information, safe from the ravages of time. Multiple, high-level studies have highlighted the critical need for and challenges of data security and long-term preservation [9, 29, 52–55]. There is even a well-regarded international standard on what an 'Open Archive Information System' needs to do [56]. Unfortunately, despite this broad understanding, most research data are likely to be lost. A recent, extensive survey of European researchers across many disciplines revealed that only 20 per cent of the respondents submit their data to an archive [31]. IPY starkly revealed this disparity between theory and practice around the world. While it is yet early for a full assessment, at the completion of IPY it appeared that only 30 of 124 large IPY science projects (24 per cent) had adequately planned for long-term preservation [57].

The IPY Data Policy required the development of project data management plans and contained strong archiving requirements, but national funding agencies are only beginning to require data management plans as a condition of funding and renewal. When they do, the requirements focus more on access than preservation. More critically, in many cases, appropriate archives simply do not exist. Many assumed that the World Data Centres, born out of IPY's most recent predecessor, the International Geophysical Year, would be natural archives for IPY data. Unfortunately, the WDCs had already been struggling and ICSU was in the process of overhauling the entire WDC system [58]. More importantly, there existed no WDC or indeed any sort of national or international archive for many of the disciplines in IPY, e.g. many social sciences, terrestrial and wildlife ecology, community health, indigenous knowledge.

IPY has made an impact, though. More funding agencies (in the US, Netherlands, UK, Canada, Norway) consider data sharing a requirement for continued support, and new archives are being established in several countries to preserve IPY data. The new ICSU World Data System (WDS) has taken on the preservation of IPY data as a major priority. The WDS promises to be a reification of the data ecosystem, but it will take time to grow and requires active involvement of major ecosystem components, notably sustained institutions and funding agencies. We hope that we trend in the right direction but recognize that there are large technical and social issues to address.

7. Discussion

We have described each of our five principles and the challenges and needs of addressing these principles with very diverse data, but ultimately the principles need to be considered in concert. Data may be open but not discoverable or useful. More commonly, data are useful but not open or safe. To truly address the challenges of very diverse, interdisciplinary data we must consider the data ecosystem as a whole. Indeed, society is still in the process of creating or evolving a fully functional data ecosystem. We, the authors, have been active players in that data ecosystem and its evolution. Between us we participate in the data collection at sea, in the field, and in the lab. We develop and present tools, information products, and scientific analyses. We curate and steward data. One of us disburses funding for data archiving and networking. We are active participants in the scientific enterprise around diverse data. In essence, our observations come from a perspective of what Latour calls 'science in action' [59]. He persuades us that it is essential to take a holistic approach to understanding social interactions and that one needs to consider the transactions of the relevant participants along with their individual actions, identities, etc. Our data ecosystem is akin to Latour's Actor-Network with its emphasis on transactions between and among human and non-human actors [60]. Data are often a focal point of these transactions and can perhaps be viewed as a key nutrient or the water that needs to flow smoothly through the ecosystem. Our goal then is to enhance that flow and enable those transactions to improve the health of the ecosystem. We also seek to foster and understand alliances and 'hybrids' of many actors [61]. We seek to not only study but also to improve the 'balance of language and practice across communities' in Star and Ruhleder's ecology of infrastructure [16].

More prosaically, our experience with a variety of IPY projects, such as the DAMOCLES project described above, illustrates how a pragmatic, flexible, and especially *collaborative* approach can encourage the behavioural adaptation in science that enhances data flow and usefulness. Collaboration between data managers and data creators increases the ecosystem transactions and balance of communication and has broad, positive repercussions. At one level, this sort of congenial collaboration can improve data and metadata completeness [51, 62]. At a deeper level, Saracevic suggested that 'the *subject knowledge view of relevance is fundamental to all other views of relevance,* because subject knowledge is fundamental to communication of knowledge' [63 p. 333, his emphasis]. In other words, we cannot communicate the value, uncertainty, or usability of data effectively without capturing expert knowledge of the data creator. Hjørland [19] argues that information science has neglected the 'subject knowledge' view of relevance and has focused unduly on a false dichotomy between 'system relevance' and 'user relevance'. Without entering that theoretical debate, as practitioners, we note that, while it is important to consider user needs and perspectives, it is equally important to consider data creator perspectives in order to effectively capture contextual knowledge. Data can serve as the sort of 'boundary objects' central for translating between viewpoints in the interdisciplinary discussion of science, but the objects must remain robust enough to maintain identity across disciplines [15]. In short, both theory and practice suggest that metadata is best generated through collaboration between the subject-knowledge holder and information specialists. This makes the data more useful within and without specialist communities.

Furthermore, Shankar illustrates how the research lab notebook becomes a significant artefact that centres, defines, and communicates knowledge creation and collaboration. [64]. Our experience echoes her view that participation in the minutiae of data creation can be a means to help form scientific identity and community. It is very helpful for data managers to participate in that process of data creation because it helps engender trust and collaboration, but it also makes the boundary object, the data, more robust. We can readily see this with reference and community data because consensus-building mechanisms such as science and technical working groups are necessarily included in their creation. With

research data there is often complete separation between the data creator and their ultimate steward (if there is one). There is a break in the flow in the data ecosystem. Communication is unbalanced and needs correction. We lack what Nardi and O'Day call a 'keystone species' in the ecosystem – the data scientist, the mediator, the translator [18].

We believe it is necessary to establish more interpersonal communication between data providers and data managers, but in many ways as the field of informatics grows more sophisticated and complex it becomes more separate from the domains it serves. This need for communication is not a new concept. Vickery [65] in his excellent review of science and technical communication in the twenty-first century notes that the need to share vast information and knowledge has always seemed daunting. The first lesson Vickery draws from his review is that 'Scientists and technologists need to interact and communicate' [65, p. 515]. We note, however, that data stewards are not solely technologists. Further, we feel that many of the proposed solutions of the past – user guides, discussion fora, etc. – while useful, do not fully address the issue. To increase interpersonal relations data managers need to participate directly in field data collection activities and scientists need to actively participate in data curation. In IPY, data management processes only began to function after a certain critical mass of interpersonal relations had been established within the science projects themselves and across the whole international data management enterprise. It is critical that scientists see themselves as part of the data ecosystem, whereas most probably do not; this change will require the time, patience, and outreach of data managers as well as flexible and useful tools for the scientists.

While we have focused on relations with the provider, it is clearly necessary to build relationships with data users as well. Not just in the sense of user-driven design but in the same collaborative sense as with providers, and again it needs to be a two way street. As Star and Ruhleder state:

> We must ask users to meet designers halfway by learning their language and developing an understanding of the design domain. If designers are at fault for assuming that all user requirements can be formally captured and codified, users are often equally at fault for expecting "magic bullets" – technical systems that will solve social or organizational problems. [16, p. 130]

This general social adaptation necessary to increase collaboration needs to continue, but it must also be supported by more sophisticated technologies. Ultimately, we seek to create and curate what might be called complex e-science objects similar to what Hunter calls a 'scientific publication package' [66]. A complex e-science object is a distributed body of diverse information potentially including physical samples, digital data, documentation, images, multi-media, publications, etc. These complex objects, not just the data, are ultimately what need to be interconnected across disciplines in the linked data network. Creation and management of these e-science objects is currently a non-standard and very manual process. Research is needed on how to automatically and reliably capture and enrich contextual information and other metadata to support reuse and archiving. A particularly challenging question that bridges both technical and social domains is how can we develop effective methods to capture data uncertainties and the tacit, domain-specific knowledge and assumptions of any data collector [67]. Increased collaboration is bound to help.

The WDS is adopting a 'data publication' model, which seeks to be somewhat analogous to publication in the literature. The idea is that published data are considered to be first-class, well-preserved, referenceable, scientific artefacts akin to journal publications [68]. Ideally, these artefacts would be the complex and dynamic e-science objects just discussed. They would be interconnected with other data, forming a combination of data publication and linked data [69], although initial work in this area suggests that this may be more complex than first envisioned. [66]. It is unclear how the extensive data publication requirements such as tracking data provenance, assuring data quality, providing fair credit and attribution, ensuring data fixity and scientific reproducibility can scale to address the huge complexity of all research data. As mentioned, research is needed on how we can automate the creation of e-science objects and persistently identify them, scientists need to be more educated on and engaged in data management from the onset of their projects, and, most critically, we need to develop the sustainable business models that can continue the preservation of rapidly growing data collections.

This need for a sustainable business model highlights the importance of sponsors and funding agencies, indeed the money itself, in the data ecosystem. In IPY, funding was both a critical gap and a significant facilitator to creating a functional data ecosystem. In their final report the IPY Data Policy and Management Subcommittee note:

> In the period leading up to the start of IPY, data stewardship was undervalued, despite robust data management plans within the IPY Framework Document, the strong recommendations of the ICSU Program Area Assessment, and telling examples from earlier international projects. [57, p. 460]

The planners were saying the right things, but early in IPY it proved extremely difficult to secure IPY data management funding, especially for broad international or interdisciplinary support. Towards the end of the IPY, more national governments began supporting national data coordinators and data archives, but international coordination and

governance remain significant challenges. Data management funding is often a problem, especially with large international programmes. We hope that IPY highlights how that needs to change and believe we are seeing some changes as a result. Sponsors are beginning to realize that, if they fund data collection, they also need to fund data stewardship. They also have a responsibility to enhance the data flow, and this is where we saw significant improvement under IPY.

As mentioned, the IPY Data Policy drove a change in several countries where research funding was contingent upon data release. This command and control approach may seem contrary to the evolutionary nature of an ecosystem, but we believe generally free flowing data are key to a healthy data ecosystem. Open data can encourage more interaction between the components of the ecosystem, and this interaction enables greater interdisciplinary understanding and collaboration. One might think of funding agency intervention as a form of ecosystem management, a controlled burn if you will, designed to bring the ecosystem in balance and more in line with stated policy. As the ecosystem evolves and more routine norms of sharing are established, less draconian methods may be employed, but it is still important to recognize sponsors as key actors in the data ecosystem.

Sponsors need to be actively engaged in the planning and execution of data systems for large experiments or campaigns. The World Ocean Circulation Experiment (WOCE) is often cited as a successful example of good data stewardship in support of a large international programme. While it was much more disciplinarily focused than IPY, it had a very long planning period. Data Assembly Centers for the different data types were funded and in place before the data collection began. The World Data Center for Oceanography in Silver Spring was the designated long-term archive. WOCE had a rigorous data policy, which required participants to submit the data to a data centre within six months of collection. This helped create an imperative to develop the data submission and access infrastructure in advance. As one senior oceanographer noted, 'Everybody involved knew the data managers, knew which data centre to submit what data to, knew which format was required, etc.'. This is in direct contrast to IPY and many other international programmes where data systems are so often an afterthought. It helped that the preparation and collaboration were driven by highly respected and well-known oceanographers who worked to move the funding agencies toward 'a different (and new) way of thinking. They did not fund "innovative research" for the usual "four-year project duration", but had to fund existing data centres and data managers, for a much longer period than four years'. This was possible partly because some of the leading programme managers on the project were respected and recognized scientists themselves, and thereby collegial participants and actors in the ecosystem. We see a similar evolution in thinking in response to IPY, but IPY and its diversity are presenting a much more complex challenge.

Moving forward, we will need to take an organic approach. We must encourage a socio-technical evolution, where the science community continually adapts to new technologies and the demands of the fourth paradigm, and the new technologies adapt to the dynamic, interdisciplinary needs of science. This creates an overarching need for flexibility and adaptability that can help us identify both technical and socio-cultural requirements or next steps toward creating a robust data ecosystem.

Technologies need to be lightweight, modular, and start simple to address complexity. We need to recognize that: (1) data will be highly distributed and housed at many different types of institutions: (2) the use and users of the data will be very diverse and even unpredictable; and (3) the types, formats, units, contexts and vocabularies of the data will continue to be very complex if not chaotic. These observations suggest some short-term technical strategies:

- Centralized registries of data and services will not fully scale to handle the diversity. Data managers need to employ more open, cloud-based approaches of data broadcasting and customized aggregation.
- Data managers need to make their data and metadata available through a variety of protocols using multiple formats to serve variable communities.
- Data systems need to start simple and iterate to expand their interconnection with other systems and user communities.
- Data system developers need to work closely with their provider community to improve acceptance and use of standards.

Over the longer term, even greater flexibility is necessary. Data systems will need to link across the large, structured, high volume systems such as the Earth System Grid; the formalized data publications of the WDS and research libraries; and the broad semantic web. Informatics research needs to explore ways to better define, describe, automatically create, and interrelate complex e-science objects across disciplines and data systems.

A larger challenge is to integrate these technical strategies within a cultural evolution. We should recognize that scientific practice is guided by broadly accepted norms of behaviour that value scientific transparency, integrity, and reproducibility. Ethically open data sharing must be a central tenet of science. Most scientific data should be viewed as a common, networked good that is generally open, except for legitimate ethical, but not proprietary, restrictions. Data

should be used in an ethical framework where data creators are given fair attribution and data uncertainties are well characterized. Data managers play a vital role in this ethical framework by working with data creators to ensure the integrity and context necessary to support robust science. Creating this ethos is clearly a long-term challenge, but again short-term strategies can help move us in the right direction:

- Data management plans and archiving needs to be required and funded as part of basic research proposals. We are encouraged that the U.S. National Science Foundation has recently mandated data management plans in proposals.[11]
- Basic data management needs to be included in the core scientific curriculum. Just like scientists of all types need to take a scientific methods class to get an advanced degree, they should also need to take a 'data class'. Many have advocated greater education of the library or data management workforce [e.g. 2, 12], but we feel it is equally important to educate the data providers.
- Data providers need to receive formal recognition of their intellectual efforts through data citation. This needs to be encouraged by funders, journals, reviewers and academia at large [38].
- Data managers need to establish close working relationships with their data providers as well as their users, based on mutual trust.

Over the longer term, data scientists need to continue to professionalize their discipline. The establishment of informatics foci and committees in international scientific societies and unions is a welcome development. These groups need to work with educational institutions, standards bodies, professional societies, and others to add increased scientific rigour and adaptability to data and systems. During IPY, we were humbled by how the rapid and dramatic changes in the polar regions increased the expectation for rigorous, scrupulously documented data. Data scientists need to be prepared for that expectation to increase dramatically. We must also do more to keep the data safe. Science and society will need to develop sustained archival business models. When we consider data a common good, this suggests that preservation of data should be a broad societal cost. It will increasingly be necessary to identify mechanisms for collaborative international funding that supports the necessarily global aspects of interdisciplinary science.

In closing, we note that the grand challenges of our society are all deeply interdisciplinary problems [70]. Some might say trans-disciplinary problems. We believe our vision of discoverable, open, linked, useful, and safe data can help address those grand challenges and that achieving this vision is a grand challenge in its own right. Considering each of these requirements specifically and in the larger context of a science data ecosystem leads us to suggest the need for a rapid socio-technical evolution. We need to recognize that informatics solutions are rapidly evolving and selection processes are going on all the time. These solutions are not usually the result of any central plan, and they require continued organic adaptation by technology, people, organizations, and society.

Acknowledgements

The authors gratefully acknowledge the intelligent insights of the reviewers who made this a more cogent paper. This material is partially based upon work supported by the National Science Foundation under grant no. 0632354, the Research Council of Norway under grant no. 18 4887/53, and by the European Union under the 6th Framework Programme.

Notes

1 https://www.teragrid.org/
2 http://polarcommons.org/
3 http://www.geoportal.org
4 http://www.opensearch.org/
5 http://sciencecommons.org/projects/publishing/open-access-data-protocol/
6 See http://linkeddata.org and http://nosql-database.org/ for more information on these concepts and technologies
7 http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
8 https://esg.llnl.gov:8443/
9 http://www.unidata.ucar.edu/software/netcdf/conventions.html
10 http://ipydis.org/data/metadata.html
11 http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf11001

References

[1] Bell G, Hey T, Szalay A. Beyond the data deluge. *Science* 2009; 323: 1297–1298.
[2] ARL Joint Task Force on Library Support for E-Science. *Agenda for Developing E-Science in Research Libraries*. http://www.arl.org/bm~doc/ARLESciencefinal.pdf (2007, accessed 27 February 2011).

[3] Hey T, Hey J. e-Science and its implications for the library community. *Library Hi Tech* 2006; 24: 515–528.
[4] Hey T, Tansley S, Tolle K (eds). *The fourth paradigm: data-intensive scientific discovery*. USA: Microsoft Research, 2009.
[5] Newman HB, Ellisman MH, Orcutt JA. Data-intensive e-science frontier research. *Communications of the ACM* 2003; 46: 68–77.
[6] Carlson DJ. Why do we have a 4th IPY? In: Barr S and Lödecke C (eds) *The history of the international polar years (IPYS)*. Berlin: Springer-Verlag, 2010.
[7] Carlson DJ. IPY 2007–2008: where the threads of the double helix and Sputnik intertwine. In: Huettmann F (ed.) *Protection of the three poles.* Springer Japan, 2011.
[8] Baker DN, Barton CE, Peterson WK, Fox P. Informatics and the 2007–2008 Electronic Geophysical Year. *Eos, Transactions of the American Geophysical Union* 2008; 89: 485–486.
[9] NSB (National Science Board). *Long-lived digital data collections: enabling research and education in the 21st century*. Washington, DC: National Science Foundation, 2005.
[10] Heidorn PB. Shedding light on the dark data in the long tail of science. *Library Trends* 2008; 57: 280–299.
[11] Hurd JM. The transformation of scientific communication: a model for 2020. *Journal of the American Society for Information Science* 2000; 51: 1279–1283.
[12] Arzberger P, Schroeder P, Beaulieu A, Bowker G, Casey K, Laaksonen L, et al. Science and government: an international framework to promote access to data. *Science* 2004; 303: 1777–1778.
[13] Doorn P, Tjalsma H. Introduction: archiving research data. *Archival science* 2007; 7: 1–20.
[14] Beers PJ, Bots PWG. Eliciting conceptual models to support interdisciplinary research. *Journal of Information Science* 2009; 35: 259.
[15] Star SL, Griesemer JR. Institutional ecology, 'translations' and boundary objects: amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science* 1989; 19: 387–420.
[16] Star SL, Ruhleder K. Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research* 1996; 7: 111.
[17] National Research Council. *Environmental data management at noaa: archiving, stewardship, and access*. Washington, DC: National Academies Press, 2007.
[18] Nardi BA, O'Day V. *Information ecologies: using technology with heart.* The MIT Press, 2000.
[19] Hjørland B. The foundation of the concept of relevance. *Journal of the American Society for Information Science and Technology* 2010; 61: 217–237.
[20] Baker J, O'Connell KM, Williamson RA. *Commercial observation satellites: at the leading edge of global transparency.* RAND Corporation. http://www.rand.org/pubs/monograph_reports/MR1229.html (2001, accessed 5 February 2011).
[21] de Sherbinin A, Chen RS. (eds). *Global spatial data and information user workshop: report of a workshop*, de Sherbinin A, Chen RS (eds) Socioeconomic Data and Applications Center, Center for International Earth Science Information Network, Columbia University. http://sedac.ciesin.columbia.edu/GSDworkshop/ (2005, accessed 5 February 2011).
[22] Esanu JM, Uhlir PF (eds). *Open access and the public domain in digital data and information for science*. Washington, DC: National Acadamies Press, 2004.
[23] Klump J, Bertelmann R, Brase J, Diepenbroek M, Grobe H, Höck H, et al. Data publication in the open access initiative. *Data Science Journal* 2006; 5: 79–83.
[24] Nelson B. Data sharing: empty archives. *Nature* 2009; 461: 160–163.
[25] OECD (Organisation for Economic Co-operation and Development). *OECD principles and guidelines for access to research data from public funding*. Paris: Organization for Economic Co-operation and Development, 2007.
[26] Schofield PN, Bubela T, Weaver T, Portilla L, Brown SD, Hancock JM, et al. Post-publication sharing of data and tools. *Nature* 2009; 461: 171–173.
[27] Lor PJ, Britz JJ. Is a knowledge society possible without freedom of access to information? *Journal of Information Science* 2007; 33: 387–397.
[28] GEO (Group on Earth Observations). *Implementation guidelines for the GEOSS data sharing principles*. Group on Earth Observations. http://www.earthobservations.org/documents/geo_vi/07_Implementation%20Guidelines%20for%20the%20GEOSS%20Data%20Sharing%20Principles%20Rev2.pdf (2009, accessed 14 January 2011).
[29] ICSU (International Council for Science). *ICSU report of the CSPR assessment panel on scientific data and information*. Paris: ICSU, 2004.
[30] WMO. *World Meteorological Organization congress, resolution 40 (Cg-XII, 1995): WMO policy and practice for the exchange of meteorological and related data and products including guidelines on relationships in commercial meteorological activities*. World Meteorological Organisation. http://www.nws.noaa.gov/im/wmocovr.htm (1995, accessed 5 February 2011).
[31] Kuipers T, van der Hoeven J. *PARSE.Insight: insight into issues of permanent access to the records of science in Europe. Survey report*. European Commission, 2009.
[32] Key Perspectives Ltd. *Data dimensions: disciplinary differences in research data sharing, reuse and long term viability*. Digital Curation Center. http://www.dcc.ac.uk/sites/default/files/SCARP%20SYNTHESIS_FINAL.pdf (2010, accessed 5 February 2011).

[33] Nicholson SW, Bennett TB. Data sharing: academic libraries and the scholarly enterprise. *Portal: Libraries and the Academy* 2011; 11: 505–516.
[34] House of Commons Science and Technology Committee. *The disclosure of climate data from the climatic research unit at the University of East Anglia, HC 387-1*. The Stationery Office Limited. http://www.publications.parliament.uk/pa/cm200910/cmselect/cmsctech/387/387i.pdf (2010, accessed 14 January 2011).
[35] Oxburgh R, Davies H, Emanuel K, Graumlich L, Hand D, Huppert H, Kelly M. *Report of the international panel set up by the University of East Anglia to examine the research of the Climatic Research Unit*. University of East Anglia. http://www.uea.ac.uk/mac/comm/media/press/CRUstatements/SAP (2010, accessed 14 January 2011).
[36] Russel M, Boulton G, Eyton D, Norton J. *The independent climate change e-mails review*. University of East Anglia. http://www.cce-review.org/pdf/FINAL%20REPORT.pdf (2010, accessed 14 January 2011).
[37] Ekborn A, Helgesen GEM, Lunde T, Tverdal A, Vollset SE. *Report from the Investigation Commission appointed by Rikshospitalet – Radiumhospitalet MC and the University of Oslo January 18, 2006*, Rikshospitalet. http://www.rikshospitalet.no/ikbViewer/Content/411234/Report%20from%20the%20Investigation%20Commission.pdf (2006, accessed 5 February 2011).
[38] Parsons MA, Duerr R, Minster JB. Data citation and peer-review. *Eos, Transactions of the American Geophysical Union* 2010; 91: 297–298.
[39] Berner ES, Moss J. Informatics challenges for the impending patient information explosion. *Journal of the American Medical Informatics Association* 2005; 12: 614–617.
[40] Cook M. Professional ethics and practice in archives and records management in a human rights context. *Journal of the Society of Archivists* 2006; 27: 1–15.
[41] Kisselburgh LG. Reconceptualizing privacy in technological realms: Theoretical frameworks for communication. *Annual meeting of the International Communication Association, TBA, Montreal, Quebec, Canada, May 21, 2008*; http://www.allacademic.-com/meta/p233000_index.html (2008, accessed 10 December 2010).
[42] Fox P, Hendler J, Semantic eScience: encoding meaning in next-generation digitally enhanced science. In: Hey T, Tansley S, Tolle K (eds) *The fourth paradigm: data-intensive scientific discovery*. USA: Microsoft Research, 2009.
[43] USGCRP. *Global change science requirements for long-term archiving*. US Global Climate Research Program, 1999.
[44] Lee CA. A framework for contextual information in digital collections. *Journal of Documentation* 2011; 67: 95–143.
[45] Best B, Halpin P, Read E, Qian A, Hazen L, Schick R. Geospatial web services within a scientific workflow: predicting marine mammal habitats in a dynamic environment. *Ecological Informatics* 2007; 2: 210–223.
[46] Bluhm B, Watts D, Huettmann F, Free database availability, metadata and the internet: an example of two high latitude components of the Census of Marine Life. In: Huettmann F, Cushman SA (eds) *Spatial complexity, informatics, and wildlife conservation*. Berlin: Springer, 2010.
[47] De Broyer C, Danis B. How many species in the Southern Ocean? Towards a dynamic inventory of the Antarctic marine species. *Deep Sea Research Part II: Topical Studies in Oceanography* 2010.
[48] Griffiths HJ. Antarctic marine biodiversity – what do we know about the distribution of life in the Southern Ocean. *PloS ONE* 2010; 5: e11683.
[49] Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 1989; 319–340.
[50] Moore R, Anderson WL. ASIS&T Research Data Access and Preservation Summit: Conference summary. *Bulletin of the American Society for Information Science and Technology* 2010; 36: 42–45.
[51] Parsons MA, Brodzik MJ, Rutter NJ. Data management for the cold land processes experiment: improving hydrological science. *Hydrological Processes* 2004; 18: 3637–3653.
[52] NSF. *It's about time: research challenges in digital archiving and long-term preservation, final report, Workshop on research challenges in digital archiving and long-term preservation.* Sponsored by the National Science Foundation, Digital Government Program and Digital Libraries Program, Directorate for Computing and Information Sciences and Engineering, and the Library of Congress, National Digital Information Infrastructure and Preservation Program. 2003.
[53] RLG/OCLC Working Group on Digital Archive Attributes, Beagrie N, Bellinger M, Dale R, Doerr M, Hedstrom M, et al. *Trusted digital repositories: attributes and responsibilities* RLG, Inc. http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf (2002, accessed 5 February 2011).
[54] Thaesis and van der Hoeven. *PARSE.Insight: insight into issues of permanent access to the records of science in Europe. Insight report.* European Commission, 2010.
[55] *The International Research on Permanent Authentic Records in Electronic Systems (InterPARES)*. School of Library, Archival & Information Studies, The University of British Columbia. http://www.interpares.org/welcome.cfm (2011, accessed 1 May 2011).
[56] ISO. *ISO standard 14721:2003, space data and information transfer systems – a reference model for an open archival information system (OAIS),* International Organization for Standardization, 2003.
[57] Parsons MA, de Bruin T, Tomlinson S, Campbell H, Godøy Ø, LeClert J, and IPY Data Policy and Management SubCommittee, The state of polar data – the IPY experience. In: Krupnik I, Allison I, Bell R, Cutler P, Hik D, López-Martínez

J, Rachold V, Sarukhanian E, and Summerhayes (eds) *Understanding earth's polar challenges: international polar year 2007–2008.* Edmonton, Canada: CCI Press, 2011.
[58] ICSU (International Council for Science). *Ad hoc strategic committee on information and data. Final report to the ICSU Committee on Scientific Planning and Review*. ICSU. http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/2123_DD_FILE_SCID_Report.pdf (2008, accessed 5 February 2011).
[59] Latour B. *Science in action: how to follow scientists and engineers through society.* Cambridge, MA: Harvard University Press, 1987.
[60] Latour B. On recalling ANT. In: Law J, Hassard J (eds) *Actor network theory and after*. Malden, MA: Blackwell, 1999.
[61] Latour B. *We have never been modern*. Cambridge, MA: Harvard University Press, 1993.
[62] Parsons MA, Duerr R. Designating user communities for scientific data: challenges and solutions. *Data Science Journal* 2005; 4: 31–38.
[63] Saracevic T. RELEVANCE: a review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science* 1975; 26: 321–343.
[64] Shankar K. Ambiguity and legitimate peripheral participation in the creation of scientific documents. *Journal of Documentation* 2009; 65: 151–165.
[65] Vickery B. A century of scientific and technical information. *Journal of Documentation* 1999; 55: 476–527.
[66] Hunter J. Scientific publication packages: A selective approach to the communication and archival of scientific output. *The International Journal of Digital Curation* 2006; 1: 33–52.
[67] Parsons MA. *Data for modelers. helping understand the climate system.* Boulder, CO: University of Colorado, 2010.
[68] Costello MJ. Motivating online publication of data. *Bioscience* 2009; 59: 418–427.
[69] Bechhofer S, Ainsworth J, Bhagat J, Buchan I, Couch P, Cruickshank D, et al. Why linked data is not enough for scientists. *Proc. of the sixth IEEE e-science conference* 2010;.
[70] ICSU. *Earth system science for global sustainability: the grand challenges*. Paris: International Council for Science, 2010.