ICSU AND THE CHALLENGES OF DATA AND INFORMATION MANAGEMENT FOR INTERNATIONAL SCIENCE

Peter Fox[1*] and Ray Harris[2]

*[*1]Rensselaer Polytechnic Institute (RPI), 110 8th St, Troy, NY 12180, United States*
Email: pfox@cs.rpi.edu
[2]University College London, Gower St, London WC1E 6BT, UK
Email: ray.harris@ucl.ac.uk

ABSTRACT

The International Council for Science (ICSU) vision explicitly recognises the value of data and information to science and particularly emphasises the urgent requirement for universal and equitable access to high quality scientific data and information. A universal public domain for scientific data and information will be transformative for both science and society. Over the last several years, two ad-hoc ICSU committees, the Strategic Committee on Information and Data (SCID) and the Strategic Coordinating Committee on Information and Data (SCCID), produced key reports that make 5 and 14 recommendations respectively aimed at improving universal and equitable access to data and information for science and providing direction for key international scientific bodies, such as the Committee on Data for Science and Technology (CODATA) as well as a newly ratified (by ICSU in 2008) formation of the World Data System. This contribution outlines the framing context for both committees based on the changed world scene for scientific data conduct in the 21st century. We include details on the relevant recommendations and important consequences for the worldwide community of data providers and consumers, ultimately leading to a conclusion, and avenues for advancement that must be carried to the many thousands of data scientists world-wide.

Keywords: Data science, Data access, Data release, Data management, Interdisciplinary science

1 INTRODUCTION

There is no doubt that scientific data and information[1] have made significant impacts on our society. The understanding of contemporary climate change is dependent upon high quality data and information; the major advances in our understanding of the origins and evolution of the universe are built upon the solid foundation of high quality astronomy data, and in the last decade, the major steps taken in understanding the human genome have been dependent on high quality data in the life sciences.

The challenges facing science as far as managing scientific data and information are concerned fall into two contrasting camps. First, there are the enormous volumes of data that are being and will be produced in science sectors such as astronomy, biomedicine, environmental science, Earth observation, and particle physics, often termed the data deluge (Hey & Trefethen, 2003), the data tsunami, or the fire hose of data. Second, there is the reluctance of some scientists to share their data because of the overheads incurred in preparing the data so that they can be shared (Nelson, 2009). Several recent reviews have implicitly or explicitly noted these two major challenges, including reports produced by the European Commission (EC, 2010; GRDI, 2011), the Organisation for Economic Cooperation and Development (OECD, 2007), and the Alliance of German Science Organisations (2008). The size of the problem can be quickly gauged in Table 1. Hilbert and Lopez (2011) have compiled estimates through modelling of the world's technological capacity to store, communicate, and compute information, and the table presents a selection of their results for storage, telecommunications, and general purpose computation. It is no surprise to see that the annual growth rates for data and data processing are very high, with the estimates for the year 2007 being far in excess of those for the year 1986. An interesting analogy on data storage is that if all the data used in the world were written to CD-ROMs and the CD-ROMs piled up in a single stack, the stack thereby created would stretch from the Earth to the Moon and a quarter of the way back again. The explosion in the quantity of data and information available to science continues apace. Whilst the absolute size of this explosion varies across disciplines, the general trend is for rapid growth in all disciplines from the social sciences to seismology, from the humanities and social sciences to high energy physics. By the end of 2011 it was estimated that 30,000 human genome sequences will have been completed (Nature, 2010b), creating information about billions of bases and requiring petabytes of data storage. A study by the International

[1] For the purposes of this paper a definition of data and information is given in Appendix A.

Data Corporation (IDC, 2010) in 2010 estimated that by the year 2020 there will be 35 zettabytes (ZB) of digital data created per annum. The IDC estimate of the total digital storage capacity in the world to be available in 2020 is 15 ZB, less than half the amount of digital data produced by then. When the Square Kilometre Array radio telescope in astronomy is fully functional in 2024, it will be able to produce more digital data than is capable of being processed in all the world's computers put together.

Table 1. Estimates of the world's capacity to store, communicate, and compute data and information. Source: Hilbert and Lopez (2011) who give detailed descriptions of the variables plus links to back-up tables for each variable.

		1986	**1993**	**2000**	**2007**	**Percent annual rate of change 1986-2007**
Storage	MB optimal compression per capita (installed capacity)	539	2,866	8,988	44,716	23
	Approximate CD-ROM equivalent per capita	<1	4	12	61	
	Percent digital	0.8	3	25	94	
Telecommunications	MB optimal compression per capita per day (effective capacity)	0.16	0.23	1.01	27	28
General purpose computation	MIPS per capita (installed capacity)	0.06	0.8	48	968	58

While the recognition of the data deluge has been relatively recent, the International Council for Science (ICSU) has been actively involved in the management of data and information since the 1950s when the World Data Centres were established as part of the International Geophysical Year of 1957-58. In its vision statement (ICSU, 2006, 2011a, 2011b) ICSU explicitly recognises the value of data and information to science:

> *[the vision of ICSU is] … a world where science is used for the benefit of all, excellence in science is valued and scientific knowledge is effectively linked to policy making. In such a world, universal and equitable access to high quality scientific data and information is a reality …*

Within the ICSU family there are organisations actively exploring how to implement the ICSU vision for universal and equitable access to high quality scientific data and information against the backdrop of the major challenges noted earlier. These organisations include the following:

- Committee on Data for Science and Technology (CODATA)
- International Network for the Availability of Scientific Publications (INASP)
- International Council for Scientific and Technical Information (ICSTI)
- World Data System (WDS)

The purpose of this paper is to explore the challenges of data and information management for international science by focussing on the ways in which the International Council for Science has reviewed and acted upon these challenges. The paper is based on discussions and reports from two *ad-hoc* ICSU committees, the Strategic Committee on Information and Data (SCID, ICSU, 2008) and the Strategic Coordinating Committee on Information and Data (SCCID, ICSU, 2011a, b). These committees produced reports that make 5 and 14 recommendations respectively aimed at improving universal and equitable access to data and information for science, and providing direction for key international scientific bodies such as CODATA and the World Data System.

2 DATA AND INFORMATION CHALLENGES

2.1 The Fourth Paradigm

The volume and complexity of data and information available to science have given rise to what some call the Fourth Paradigm of science (Hey et al., 2009). This fourth paradigm puts data-intensive science into the context of its three main predecessors, namely (Bell et al., 2009):

- First Paradigm. Observation, descriptions of natural phenomena, and experimentation.
- Second Paradigm. Theoretical science such as Newton's laws of motion and Maxwell's equations.
- Third Paradigm. Simulation and modelling, such as in astronomy.
- Fourth Paradigm. Data-intensive science that exploits the large volumes of data in new ways for scientific exploration, such as the International Virtual Observatory Alliance in astronomy.

The Fourth Paradigm is not only characterised by massive data volumes but also by complexity of data sets and by the potential for extensive cross-fertilisation of data, information, information technology, and publishing. The Fourth Paradigm acknowledges the central role played by data in science and in some ways reflects the empirical but computationally-limited First Paradigm.

2.2 Data overload

The last decade has seen substantial change in the creation, use, and management of scientific data and information, not least amongst scientists, data managers, libraries, and publishers. The traditional reward mechanisms for scientists have been in grants, publications, citations, prizes, and promotion. There is now a strong interest in publishing data and for such publication reward or recognition systems do not commonly exist. When scientists use data they must often now be concerned with the conditions of access to the data, for example, copyright, onward distribution, and use licences, as well as with the data themselves, and they must also enter the arena of standards and interoperability so that they can read the digital data needed for their work and produce outputs that are accessible to other scientists. Data managers are now often in charge of very large data repositories, for example in astronomy, and they need to provide tools to help scientists use data.

In the last decade there has been a rapid expansion of the responsibilities of libraries to encompass digital repositories, including data repositories, alongside traditional books and journals. This means in particular that there is a need for knowledge of deposit and access conditions, digital rights such as Creative Commons licences, and the use of standards, metadata schemes, and persistent identifiers, such as those promoted by DataCite (2011) to ensure correct data citation. In parallel, publishers have also made major changes to encompass digital data. Some publishers encourage, or even require, the submission of data to either their own journals as supplemental material or to recommended data centres. As an illustration, the journal *Nature* makes it mandatory for certain types of human genome data that are associated with accepted publications to be submitted to a community-endorsed, public repository: for example, DNA and RNA sequence data have to be submitted to the Protein DataBank, UniProt, or GenBank/EMBL/DDBJ nucleotide sequence database.

Open access journals have been changing the landscape of journal publishing away from the traditional model of payment by subscribers and libraries. More than 6,000 titles are currently registered in the Directory of Open Access Journals. Traditional publishers are experimenting with new business models and increasingly offering open access options, for example the relationship between the publisher Elsevier and the PANGAEA data centre in Germany.

2.3 Data complexity

There are four important characteristics of complex data: high dimensionality, multimodality, multi-scale, and heterogeneity. Multimode data appears in fields ranging from neuroscience to astronomy, and while its origins are in imagery, it is now appearing in application areas such as air quality where the modes of measurement are very different. In a range of fields from environment and climate to biomedicine, crossing scales has emerged as a key need, for example, crossing the scales from molecular to cell to tissue and then to organ and organism scales in animals, each of which has different measurement and structural data representations. While there are promising approaches to reduce complexity, further complications such as dependency among dimensions may

result in redundancy and inaccuracy in semantics. However, progress using a variety of means (algorithmic, representational, and computational) is beginning to occur in some fields. In the present context, this is all dependent on data and information and the application of data science.

2.4 Changing expectations

Expectations on scientists in the area of data and information management have evolved and increased over the past decades as science itself has moved into the data-intensive era. The main drivers of these changing expectations are the changing nature of science, science funders, policy makers, and governments as well as society at large. Science is more than ever a globalized international activity with a strong collaborative component. To carry out their research, scientists are not only expected to manage, share, and archive their data professionally but also to use cutting-edge information and communication technologies for data and information discovery and analysis. Unfortunately, the vast majority of scientists who work with data are neither well equipped nor trained to meet these high expectations. On the other hand, data scientists are working at the forefront of information technology and have the knowledge to develop the tools and training in this important area of data management.

Scientists have to respond and adapt to new expectations coming from governments and funding agencies, such as the National Science Foundation in the United States, which are increasingly requiring a full data management plan to be submitted with applications for research funding. Scientists are also facing new expectations from society at large as the outcome of their research is used by policy makers in designing public policies that affect society directly and by applied users from both the public and private sectors. Scientists need not only to communicate honestly and openly their research but also to share and open their data to public use and media scrutiny, as illustrated in the field of climate change by the so-called Climategate scandal (Nature, 2010a).

2.5 Digital divide

While there are rapid advances in data capture technologies and the ability to handle the data deluge, there is still a digital divide with those scientists in the less economically developed countries (LEDCs) who lack access to both data and technology. The meetings of the World Summit on the Information Society in both 2003 and 2005 identified the digital divide as a major concern for society. Data on computer availability (UN, 2008) show that while the countries in the North have better than one computer for every two people, the LEDC countries have about one computer for every 10-20 people. Broadband penetration in LEDCs lags similarly behind the provision of computers although undersea cables are set to have a major impact on connectivity in African countries. Data from the International Telecommunication Union for 2009 (ITU, 2009) show that there is only one fixed broadband subscriber for every 1,000 people in Africa compared to one for every 200 in Europe. In the countries of the North, National Research and Education Networks (NRENs), such as GEANT2, SINET, and AARNet, have developed alongside commercial broadband capacity to provide dedicated services and support to research and education. NRENs are either absent from or only recently emerging in the LEDCs, particularly in Sub-Saharan Africa. Whilst there have been significant recent connectivity developments in South Africa and Kenya, the picture in the rest of Africa is still very much one of limited or poor connectivity.

3 ICSU STRATEGIC REVIEWS OF INFORMATION AND DATA

During the last decade, ICSU has launched several strategic reviews of the capability of international science to handle the growing volume and complexity of data and information. In 2004 ICSU's Panel Area Assessment (PAA) on Information and Data (ICSU, 2004) identified three main requirements for improved data and information management in science. First, to ensure universal and equitable access to data and information. Better access will lead to better science. Second, to develop an improved capability to manage data professionally, vital both for access to good quality data now and to ensure that future scientists will have access to historical data. Third, to consider the question of who pays for data and for professional data management because reliable funding is always required for the creation and management of data and information: no funding means no data. The PAA report was extensive in its recommendations but kept returning to ways to encourage and enable the scientific community to improve its strategic capability to think about and then take action on data and information management, both within the ICSU family of national members and scientific unions and in relation to other organisations.

Following the PAA report, ICSU established a Strategic Committee on Information and Data (SCID) to examine how in practice to facilitate a new, coordinated global approach to scientific data and information that ensures equitable access to quality data and information for research, education, and informed decision-making. The SCID report (ICSU, 2008) recommended the creation of a new World Data System (WDS) based on the former World Data Centres (WDCs) and the Federation of Astronomical and Geophysical data analysis Services (FAGS). The purpose of the World Data System is to provide a coordinated, professional approach to the management of scientific data and the production of services based on the data. The World Data System is described in the next section of this paper. In addition, the SCID report encouraged CODATA to become more prominent in science by having clearer strategic goals that are linked with ICSU's vision for science. ICSU has national members and scientific union members, and these members provide a vital means of communication throughout the scientific community. ICSU members were also encouraged by the SCID report to engage proactively in professional data management for science.

The most recent ICSU examination of strategy for scientific data and information management has been the Strategic Coordinating Committee on Information and Data (SCCID, ICSU, 2011). SCCID continued the process of developing strategic priorities for the improvement of professional data management in science.

4 PROGRESS WITH THE WORLD DATA SYSTEM

4.1 WDS objectives from the ICSU perspective

In exploring desirable attributes of an 'ideal' system, the SCID process included a separation of these attributes into three categories: Mission, Coordination, and Execution (SCID Report, pp. 13-14.). The aim of this distinction was to match certain functions with existing international organizations (e.g., ICSU, CODATA) but importantly to identify functionality gaps in both national and international (i.e., world) data centres. In essence, an ideal system became a combination of existing activities as well as the new World Data System. For example, for the Mission of an ideal system, a) Enable and encourage the advancement of science through the open provision of high quality data and information services, b) Increase global knowledge and reduce the knowledge divide between richer and poorer countries by providing universal and equitable access to scientific data and products, c) Identify structural gaps in data and information provision and seek solutions to fill these gaps, and d) Develop further the structure for long term stewardship of scientific data, including in the form of formal public libraries for data. Coordination included: a) Fostering multi-disciplinary, large scale, complex science, b) Leading and championing professional data management, and c) Informing discussions on data policy from a science perspective. Finally Execution emphasized: a) Taking a lead role in developing, testing, and implementing standards for data access to provide services for all scientists, b) Promoting the publication of data and data products, with the associated recognition and accreditation that are common to peer-reviewed science publications, c) Providing reliable and trustworthy science-reviewed data and derived products, d) Serving discipline-based science communities with exemplary data repositories and data products, e) Integrating data sets using community-consensus algorithms, and lastly, f) Enabling seamless access to data.

The WDS adopted the vision provided in the SCID (2008) report and has articulated the following goals (Minster et al., this volume, http://www.icsu-wds.org/): a) Enable universal and equitable access to scientific data and information, b) Facilitate better access to data, c) Strive for simpler access to data, d) Work to provide quality assured data and information, e) Promote improved data stewardship, f) Work to reduce the digital divide, and g) Ultimately, provide data for better science. These goals represent an initial amalgam of SCID's Mission and Execution suggestions with a focus on near and medium term activities.

4.2 Criteria for professional data management in the WDS

The SCCID report (ICSU, 2011) took to task the issue for professionalization, both for data science and data management, and the task for the WDS is challenging. In particular, SCCID in their 'Recommendation 6' stated: *We recommend the development of education at university and college level in the new and vital field of data science. The example curriculum included in appendix D [of the SCCID report] can be used as a starting point for course development.* Also, in Recommendation 7: *We recommend that both the CODATA and the World Data System biennial conferences include forums for data professionals, including data librarians, to share experiences across a range of science disciplines.*

WDS has an opportunity to share experience with the broader community on their experience with *what are the required credentials, knowledge and skills (technical, scientific, personal, user needs, etc.) to train and give data professionals more explicit recognition* (SCCID report, p. 21). The WDS must participate in the identification and definition of a community of data professional peers and provide a variety of fora for them to meet and exchange ideas, experiences, and solutions.

4.3 Active participation in the WDS

The WDS, via its Scientific Committee (WDS-SC), is responsible for soliciting active participation in the WDS. The forms of membership include: Regular, Network, Partner, and Associate. At the time of writing there were 29 Regular members and 1 for each of the remaining categories (see: http://www.icsu-wds.org/organization/structure/wds-sc for current statistics). As a mark of the new diversity in the WDS, the active member list comprises: past World Data Centres (WDCs), Federation of Astronomical and Geophysical data analysis Services (FAGS), research institutes, international scientific unions, consortia, and commercial publishers, all of which bring a welcome diversity to the WDS. The future seems bright for a revitalized and synergistic World Data System.

4.4 Future plans for the WDS

The WDS-Transition Team (WDS-TT) and subsequently the WDS-SC took responsibility for SCID's second recommendation to 'work closely with CODATA and with the new ICSU ad hoc Strategic Coordinating Committee.' In particular, the WDS-SC efforts to develop and begin implementation of a strategic plan for the WDS will be key to its future. Now that the WDS is forming and the SCCID term has ended, additional consideration must be given to exactly which entities, beyond CODATA, the WDS-SC, and the WDS, to work closely with as well as how they interact. For example, the structure for making scientific data and information management effective and efficient will need to be re-conceived because the SCCID role of coordination has ended and the interaction and coordination roles need to be re-defined. SCCID recommendations 2 (open access), 4 (beyond the science community), and 7 (new forums for data professionals) are significant for ICSU and require adequate resource allocation and attention for the desired outcomes to be achieved. SCCID recommendation 8 (explicit visibility enhancement of data and science engagement) appears well within the current WDS-SC strategy and commensurate with the emerging set of WDS members. However, the later SCCID recommendations including 10 (exploitation of standards expertise), 12 (multi-way organizational engagement for closing the digital divide), and 14 (in-reach, raising the profile of data science) are expected to place a strain on the current WDS capability and capacity to participate beyond what many institutional hosts for World Data System nodes may consider their core mission. Experience from both SCID and SCCID deliberations suggest that deliberate and frequent communication between ICSU, WDS-SC, CODATA, other relevant ICSU inter-disciplinary bodies, International Scientific Unions, and ICSU National Members will be required.

5 FUTURE IMPROVEMENTS IN DATA AND INFORMATION MANAGEMENT FOR SCIENCE

As noted earlier, after the production of the SCID report and the stimulation for the creation of the World Data System, ICSU took a wider view of other data and information challenges in its Strategic Coordinating Committee on Information and Data (SCCID). The purpose of this section of the paper is to present some of the key recommendations of the SCCID report.

5.1 Best practice in data management

Advice and guidance on the principles of best practice in data and information management is needed, both for the members of the ICSU family and for all of science. Every sector of science can learn from previous experience in professional data management, and improvements in data management will lead to better science by improving access to data and information. A short, practical guide to best practice for professional data management in science has been produced and it is included in Appendix B of this paper. The guide draws on experience from (amongst others) the Protein Data Bank, the International Polar Year, the Intergovernmental Panel on Climate Change, and the International Virtual Observatory Alliance in astronomy.

5.2 Open access

The Open Access movement emerged in the new era of electronic information, and the concept was initially introduced and formalised in the field of access to publications through the "3B" declarations listed below.

- Budapest Open Access Initiative: http://www.soros.org/openaccess/read.shtml
- Bethesda Statement on Open Access Publishing: http://www.earlham.edu/~peters/fos/bethesda.htm
- Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities: http://oa.mpg.de/lang/en-uk/berlin-prozess/berliner-erklarung/

The 34 members of the Organisation for Economic Co-operation and Development (OECD) have agreed at ministerial level to a statement on *OECD Guidelines and Principles for Access to Research Data from Public Funding* (OECD 2007). On open access the OECD principles state:

> *Openness means access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination. Open access to research data from public funding should be easy, timely, user-friendly, and preferably Internet-based.*

The OECD principles cover 13 topics in total, including transparency, legal conformity, interoperability, and quality. The principles are regarded by the OECD as "soft law", that is they have a moral authority and strong support by ministers but they are not legally binding on OECD member states.

The open access notion has been extended to various degrees of unlimited access to online data and information and, in the domain of research data, is clearly related to the needs and practices of data sharing and re-use. As open access has a generally positive impact on scientific progress, it is increasingly supported, via formal statements and policies, by research institutions, scientific unions, government bodies, and funding agencies. However, the terminology used in open access is uncertain and at times confusing. Uncertainty has been created by the use of different ideas such as full and open access, free access, public access, universal, and equitable access and by the (somewhat artificial) distinction between access to data and access to publications. At the same time some initiatives have been trying to formalise 'open' beyond the initial access definitions, for example, open data, open archives, open content, open knowledge, and open notebook science. There is certainly merit in establishing a forum for the exploration and eventual agreement in relation to science of all the terms used under the broad umbrella of Open Access. Without agreement it seems likely that uncertainty will grow.

5.3 Data as a publication

The evolution of data analysis and the publication of scientific research are parallel to the development of science paradigms described earlier in this paper. In the first to third paradigms, data were contained within scientific papers in that the results from scientific theories, models, and experiments were presented within the paper, at least in summary form. The fourth paradigm (Hey et al., 2009) implicitly or explicitly disconnects the results of research from the data that were used to prepare the research findings, in that the data are too voluminous to publish in a conventional form. This evolution, closely linked both to scientific progress and technological advances, calls for a fresh view of the concept of "publishing" data sets in trustworthy repositories with long-term sustainability prospects. The concomitant recognition of and credit accorded to such activities are viewed increasingly as essential to such endeavours. As a result, the roles of libraries and publishers are changing in regard to data and information. Most countries have a national deposit library, which has a legal responsibility to hold a copy of any publication, and such legal deposit libraries might provide a valuable vehicle for ensuring long-term data stewardship. As an example, the national deposit library in The Netherlands is one of the world leaders on considering data as a publication and so requiring easy access and long term stewardship. At the same time, methods for citing data have evolved rapidly, such as the Digital Object Identifier concept. This provides a ready link with peer-recognition of the work of scientists who produce high quality data and with the many emerging mechanisms to communicate science.

5.4 The role of education

The conduct of science research is increasingly data driven: from data assimilation through modelling, simulation, and visualisation to long-term time series of data. It is now well established that data have an intrinsic value that outlasts current science foci. Unfortunately there is only part time attention given by most scientists to data science. Perhaps most important is the need to give a new value to science in the form of data citation, attribution, and data publication. It is essential to identify the required credentials, knowledge, and skills (technical, scientific, personal, user needs, etc.) to acquire and to give data professionals more explicit recognition. Formal training in the key cognitive and skill areas of data science will enable graduates to become key participants in eScience collaborations. The need is to teach key methodologies in application areas based on real research experience and build a skill-set.

6 CONCLUSIONS

6.1 Momentum

In science and the popular press today, barely a week goes by without an article, blog, or social media dissemination appearing, focusing on 'data'. Big data is hot news: "floods", "tsunamis", "tidal waves", exascale, etc., but so are other elements of complex data such as dimensionality, scale, modality, source heterogeneity, and inter-disciplinarity. Importantly, these factors are now being placed in plain view of the scientific community and the responses are not just coming from science but from funding and operational agencies, governments, commercial sectors, and the private sector. Some truly inspiring opportunities lie ahead. However, it is very important to point out that while many consider this new news, ICSU has been paying very careful (though perhaps without being highly visible) attention since well before 2004 when the PAA (on Information and Data) report was released. The visibility was substantially increased with the SCID and SCCID activities, which substantially took on the role of strategic examination of sectors, and needs to fulfil the ICSU vision, as well as stimulate active coordination among key organizations. Around 2005-2006, several "International Year" activities were being planned as 50th year celebrations of the International Geophysical year (1957-1958). The Electronic Geophysical Year, the International Polar Year, the International Heliospheric Year, and the International Year of Planet Earth ran approximately from 2007-2008, and all had data at the forefront. Of note was the link back to IGY and further back to previous IPYs and that for all our modern advances and technology, effect management and use of data was still a tremendous challenge. Looking back, it was most likely the confluence of national and international attention with these celebratory IGY community efforts that truly allowed data science and informatics to fully emerge in their respective areas and reinforce that ICSU was clearly paying attention and stepping up for its role in strategic coordination.

As an aggregate, the early 'results' arising out of much greater awareness among international organizations around data, willingness to broaden the conversation, and a much more inclusive trend are very promising in advancing the ICSU vision for data and information. Exemplars include the last three ICSTI conferences devoted to advanced aspects of large data and visualization, the formation of a CODATA Task Group on Data Publication, the aforementioned WDS membership composition and response, and leading efforts such as the Polar Information Commons (PIC: http://www.polarcommons.org/) and their PIC 'badging' efforts are truly advancing the discussion around data as a first class science object and in turn challenging more traditional approaches to data.

It is not possible, however, to present a uniformly glowing report of responses. Both the SCID and SCCID reports strongly emphasized the role for ICSU national members and ICSU Scientific Unions as being essential stakeholders and participants in a cohesive future. After all, the universality of science lies at the heart of the ICSU vision, and the direct resources and the attention of ICSU national members are required to implement such a vision. Several nations, or aggregates of nations, as well as scientific unions have provided substantial (and in some cases, long standing) responses to the presently articulated data agendas. Unfortunately, as a whole, the response of nations and scientific unions is poor and remains a challenge for ICSU but more-so for the scientific communities. Such under-served outcomes may ultimately lead to more and undesirable digital divides.

On the matter of extant digital divides, the EDC-LEDC distinction and the need to erase or at least dramatically reduce the inherent disadvantages faced in LEDCs in the contemporary digital information world, opens up significant opportunities for scientists with minimal resources (data, computation) to become data scientists and

thus be thrust well into a small but growing cadre of such career professionals. The opportunity is also fraught with cultural challenges but is clearly on the agenda for organizations such as the WDS and CODATA.

As we bring this paper to a close, one trend in the levels of discussions and participation reported herein is notable. ICSU activities such the PAA, SCID, SCCID, WDS-SC involve approximately tens of people, and this is true for executive/ committee leads for related organizations; ICSTI, INASP, PIC, etc. Their greater activities, such as conferences and workshops, in turn reach often hundreds of participants (e.g., the WDS Science Conference featured in this volume). These numbers fall far short of the much greater audience penetration that is needed. This means penetration through to scientific unions, professional societies, and the working ranks where the numbers are in the thousands. In other words, scaling is needed to move from tens to hundreds to thousands; otherwise the universality of science and the data and information that underpins it will be incomplete.

6.2 Avenues for action

In conclusion, we see several avenues for action for a variety of stakeholders.

We call on present and prospective new WDS members of all types to be forward looking and conversant with the greater goals and vision embraced by ICSU on behalf of the entire scientific community. The extant reports articulate many attributes and details of these goals, including new roles and new partners.

New communities are entering the conversation regarding data. Cultural, organizational, and economic barriers for publishers, librarians, and technical solutions from commercial sectors are ever present. We suggest that the required ensuing conversations and the explorations and demonstrations of mutual benefit, often measured by very different means, are an initial step worth pursuing.

Coordination of inevitably overlapping roles and responsibilities (and often authorities) around data, information and its generation, access, and use is essential. The tendency to ignore the reality that true coordination and collaboration are actually carried out among *individuals* in organizations is yet one more barrier to progress (especially noting that an individual may participate in many organizations). Even so, while the coordination opportunities abound, often the resources and attention assigned to them are discordant. We encourage that strategic attention in each organization be paid to timely and valuable coordination activities, retaining a level of agility to respond to new and changing needs.

To begin the immense task of increasing participation, or in turn addressing the scale of penetration of science communities' attention to data, an immediate avenue is effective engagement at the professional society and scientific union level. Motivated and knowledgeable scientists and managers can introduce discussions of data policies, access, management, etc. at any and every turn. A new group of peers is emerging: those with a career approach to data and information.

Data scientists are here to stay, but their explicit numbers are small as are the programs for educational preparation. A clear call to action is the introduction of initially graduate level courses, leading to the wide spread establishment of data science curricula, and degree and career paths for data scientists. In the longer term, undergraduate majors in data science are inevitable. The future will tell the remainder of the story.

7 ACKNOWLEDGEMENTS

The authors were both members of the ICSU Strategic Committee for Information and Data (SCID), 2007 - 2008 and the Strategic Coordinating Committee for Information and Data (SCCID), 2009 – 2011. We are very grateful to all our colleagues on the two committees for the extensive and detailed discussions that have informed this paper. The SCID and SCCID reports including the members of the committees can be accessed at the ICSU web site: www.icsu.org.

8 REFERENCES

Alliance of German Science Organisations (2008) *Priority Initiative "Digital Information"*, Retrieved February 2012 from WWW:
http://www.dfg.de/download/pdf/foerderung/programme/lis/allianz_initiative_digital_information_en.pdf

Bell G, Hey, T., & Szalay, A. (2009) Beyond the data deluge, *Science* 323, 6 March 2009, pp. 1297-1298.

DataCite (2011) *DataCite*. Retreived from the WWW, November 21, 2012: http//:www.datacite.org.

EC (2010) *Riding the wave. How Europe can gain from the rising tide of scientific data*, *Final report of the High Level Expert Group on ScientificData,* Brussels: European Commission

GRDI (2011) *Global Research Data Infrastructures: The GRDI2020 Vision*, project funded by the European Commission, 7th Framework Programme for Research and Technological Development. Retrieved from the WWW, November 21, 2012: http://www.grdi2020.eu/

Hey T, Tansley, S., & Tolle, K. (2009) *The Fourth Paradigm. Data-intensive scientific discovery*, Redmond, Washington: Microsoft Research

Hey, A. J. G., & Trefethen, A.E (2003) The data deluge: an e-science perspective, in F Berman, G Fox and A J G Hey (eds) *Grid Computing - Making the Global Infrastructure a Reality*, Chichester, UK: Wiley and Sons

Hilbert, M. & Lopez, P. (2011) The world's technological capacity to store, communicate and compute information, *Science* 332, 1 April 2011, pp. 60-65.

ICSU (2004) *Panel Area Assessment on Information and Data*, Paris: International Council for Science

ICSU (2006) *ICSU Strategic Plan 2006-2011*, Paris, International Council for Science

ICSU (2008) *Ad hoc Strategic Committee on Information and Data*, *Final Report to the ICSU Committee on Scientific Planning and Review*, Paris: International Council for Science

ICSU (2011a) *Ad hoc Strategic Coordinating Committee on Information and Data*, *Final Report to the ICSU Committee on Scientific Planning and Review*, Paris: International Council for Science

ICSU (2011b) *ICSU Strategic Plan II, 2012-2017*, Paris: International Council for Science

IDC (2010) *IDC Digital Universe Study*, sponsored by EMC, May 2010. Retrieved from the WWW, November, 21, 2012: http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm

ITU (2009) *The World in 2009, ICT facts and figures.* http://www.itu.int/ITU-D/ict/material/Telecom09_flyer.pdf

Nature (2010a) Closing the Climategate, *Nature* 468, 18 November 2010, p. 345.

Nature (2010b) Genomes by the thousand, *Nature* 476, 28 October 2010, pp. 1026-1027.

Nelson B (2009) Empty archives, *Nature* 461, 10 September 2009, pp. 160-163.

OECD (2007) OECD Principles and Guidelines for Access to Research Data from Public Funding, http://www.oecd.org/science/scienceandtechnologypolicy/38500813.pdf

UN (2008) *United Nations Global Development Goals Indicators 2008*. Retrieved from the WWW, November 21, 2012: http://mdgs.un.org/unsd/mdg/Default.aspx

9 Appendix A

Definition of data and information

Data and information can be considered as a continuum ranging from raw research data through to published papers. "Data" includes, at a minimum, digital observation, scientific monitoring, data from sensors, metadata, model output and scenarios, qualitative or observed behavioural data, visualizations, and statistical data collected for administrative or commercial purposes. Data are generally viewed as input to the research process.

"Information" generally refers to conclusions obtained from analysis of data and the results of research. But the distinction between data and information is flexible and will vary according to the situation. Increasingly, the output of research (traditionally viewed as "information") includes data and has become input to other research, rendering the output-input distinction between data and information meaningless.

10 Appendix B

Principles of best practice for data and information management

1. **Policy**

   - Document early the reason(s) for the data policy and the policy itself, and make documents available online.
   - Articulate the desired outcomes of the data policy.
   - Identify and be explicit about the benefit/cost ratio of professional data management.
   - Ensure that guidelines for participation are easily accessible by encouraging open access to data policies, practices, and experiences.

   Examples

   - ICSU World Data System data policy, available at: http://www.icsu-wds.org/organization/data-policy
   - International Polar Year data policy, available at: http://classic.ipy.org/Subcommittees/final_ipy_data_policy.pdf
   - OECD Principles and Guidelines for Access to Research Data from Public Funding, 2007, available at: http://www.oecd.org/science/scienceandtechnologypolicy/38500813.pdf
   - Panton Principles for open data in science, see: http://pantonprinciples.org/
   - Creative Commons licences, available at: http://creativecommons.org/choose/

2. **Governance**

   - Ensure that data management is an integral and funded part of project planning, approval and performance measurement.
   - Appoint expert advisory groups where necessary, and charge them with defined tasks.
   - Exploit major international science conferences and events as dates/locations to hold meetings, and use these meetings to encourage interactions between scientists and data/information professionals.
   - Acknowledge the different skills and roles required in professional data and information management.
   - Ensure open, online access to all minutes of meetings and decisions taken.

   Examples

   - The core agreement for the Worldwide Protein Data Bank, 2003, available at: http://www.wwpdb.org/wwpdb_charter.html
   - The Intergovernmental Panel on Climate Change structure and working groups, see: http://www.ipcc.ch/working_groups/working_groups.htm

3. **Planning and organisation**

   - Consider the advantages and disadvantages of distributed versus centralised data repository models in the light of user needs.
   - Use service-based data access methods.
   - Exploit what already exists for data management.
   - Data infrastructure should be completed, ready, and available in time for its use by scientists in research projects.
   - Incorporate user feedback into all aspects of the data management lifecycle.

Example

- GenBank, the annotated collection of all publicly available DNA sequences, see: http://www.ncbi.nlm.nih.gov/genbank/GenbankOverview.html

4. Standards and tools

- Use international standards (e.g., ISO, OGC, XML, GML) where possible, and if not possible then base domain-specific standards on international standards.
- Provide tools to support the implementation of the standards used, including documentation on how to use the project data.

Examples

- Dublin Core Metadata Initiative, available at: http://dublincore.org/documents/dces/
- ISO 19115 for geographical information and services, available at: http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020
- Open Geospatial Consortium standards and specifications, see: http://www.opengeospatial.org/standards
- International Virtual Observatory Alliance, documents and standards, available at: http://www.ivoa.net/Documents/

5. Data management and stewardship

- Minimise uncertainty at all phases of the data lifecycle, including, for example, working with manufacturers to avoid device dependency for data and information.
- Embrace science-programme and project-level data management planning.
- Ensure that documented plans for long term stewardship of data exist.
- Implement a plan for formal process for data and information selection and appraisal.
- Produce a plan for data stewardship at the outset of a project or programme, not as the last item in the plan.

Examples

- International Polar Year Data and Information Service.(has evolved to the http://www.polarcommons.org/)
- Research Information Network, stewardship of digital research data – principles and guidelines, 2008, at: http://www.rin.ac.uk/our-work/data-management-and-curation/stewardship-digital-research-data-principles-and-guidelines

6. Data access

- Minimise the burden on the providers of data.
- Provide a single portal for user discovery from distributed sources of information.
- Implement open access policies where appropriate.

Examples

- GEO portal, see: http://www.geoportal.org/web/guest/geo_home
- Ocean Data Portal, see: http://www.oceandataportal.org/

(Article history: Available online 19 January 2013)