Ad-hoc Strategic Coordinating Committee on Information and Data

Interim Report to the ICSU Committee on Scientific Planning and Review

ICSU
International Council for Science

ICSU – International Council for Science

Founded in 1931, the International Council for Science (ICSU) is a non-governmental organization representing a global membership that includes both national scientific bodies (121 National Members representing 141 countries) and International Scientific Unions (30 Members). The ICSU 'family' also includes more than 20 Interdisciplinary Bodies—international scientific networks established to address specific areas of investigation. Through this international network, ICSU coordinates interdisciplinary research to address major issues of relevance to both science and society. In addition, the Council actively advocates for freedom in the conduct of science, promotes equitable access to scientific data and information, and facilitates science education and capacity building. [www.icsu.org]

ISBN 978-0-930357-85-6

Cover image: Network cables and servers in a technology data center / istockphoto

Ad-hoc Strategic Coordinating Committee on Information and Data

Interim Report to the ICSU Committee
on Scientific Planning and Review

April 2011

Contents

Glossary of terms

AARNet	Australia's Academic and Research Network
CFRS	Committee on Freedom and Responsibility in the conduct of Science
CODATA	Committee on Data for Science and Technology
CONICYT	Science and Technology Research Council of Chile
FAGS	Federation of Astronomical and Geophysical Data Services
GEANT2	High bandwidth academic network serving Europe's research and education community
GEO	Group on Earth Observation
GEOSS	Global Earth Observation System of Systems
GML	Geography Markup Language
ICSTI	International Council for Scientific and Technical Information
ICSU	International Council for Science
IEEE	Institute of Electrical and Electronics Engineers
INASP	International Network for the Availability of Scientific Publications
IODE	International Oceanographic Data and information Exchange
IPO	International Programme Office
IPY	International Polar Year
ISO	International Standards Organisation
IVOA	International Virtual Observatory Alliance
LEDC	Less economically developed country
NREN	National Research and Education Network
ODINAFRICA	Ocean Data and Information Network of Africa
OECD	Organisation for Economic Co-operation and Development
OGC	Open Geospatial Consortium
PAA	Priority Area Assessment
REN	Research and Education Network
SADC	Southern African Development Community
SCCID	Strategic Coordinating Committee on Information and Data
SCID	Strategic Committee on Information and Data
SINET	Science Information Network of Japan
UN	United Nations
WDC	World Data Centre
WDS	World Data System
WSIS	World Summit on the Information Society
XML	Extensible Markup Language

1. Executive Summary

The 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.

The Strategic Coordinating Committee on Information and Data has produced an interim report that makes 14 recommendations to improve universal and equitable access to data and information for science. These 14 recommendations are presented below in summary form.

1. ICSU should ensure that National Members and Union Members adopt the guide to best practice presented in Appendix B of this report, either through their own data and information committees or commissions (where these exist), or independently. ICSU should also ensure that the guide is followed by all new ICSU projects and programmes.
2. ICSU should establish a forum for the exploration and eventual agreement in relation to science of all the terms used under the broad umbrella of Open Access.
3. ICSU should use the OECD guidelines that have already been agreed implicitly by 33 of its National Members, and have provided a general framework for several existing discipline-specific data access and sharing policies, as the basis for a forum to discuss and agree a set of principles among all ICSU National Members.
4. ICSU should engage actively with publishers of all kinds together with the library community and with scientific researchers to document and promote community best practice in the handling of supplemental material, publication of data and appropriate data citation. The WDS conference to be held in Kyoto in September 2011 provides a very convenient starting point for this engagement.
5. CODATA should consider as the theme for its 2012 biennial conference how data science can support the delivery of the science goals of the major ICSU Earth System Research for Global Sustainability initiative and the Planet under Pressure conference organized by ICSU's Global Environmental Change programmes planned for March 2012 in London.
6. We recommend the development of education at university level in the new and vital field of data science, using the curriculum included in an appendix D of this report as a starting point.
7. Both the CODATA and the World Data System biennial conferences should include forums for data professionals to share experiences across a range of science disciplines.
8. The WDS, once fully established, should increase the visibility of data centres and their data management procedures within the scientific community.
9. We recommend the analysis of storage models and means, including the possibility of creating an analogue with digital deposit libraries.
10. ICSU should exploit more fully the expertise in data standards already present in CODATA, the WDS and in its Scientific Union Members to assist in the definition and maintenance of high level data standards appropriate to meet both disciplinary requirements and overall science interoperability standards.
11. ICSU should develop a better mechanism to insert a science perspective into general standards bodies such as ISO, OGC, IEEE and the World Wide Web Consortium. Suitable expertise exists in the ICSU family but it is scattered in an uncoordinated way across Scientific Unions.
12. We recommend that ICSU uses CODATA, the WDS and the National and Union Members in a coordinated way to improve access to data and information in less economically developed countries.
13. ICSTI should enlarge its existing dialogue with the private sector to include both more commercial companies and more ICSU National and Union Members to explore how science and commerce can exploit data and information to mutual benefit.
14. The WDS should be the natural home for science in-reach activities and should work with CODATA on raising visibility of data and information management by scientists.

2. Context

2.1. The ICSU vision

There is no question that scientific data and information 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 ICSU vision[1] 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:

> *The long-term ICSU vision is for 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 and all countries have the scientific capacity to use these and to contribute to generating the new knowledge that is necessary to establish their own development pathways in a sustainable manner.*

ICSU has for many years promoted the freedom of communication, movement and association of scientists so that the free flow of ideas and theories is encouraged and facilitated. The same can be said for data and information, and in its *Strategic Plan 2006-2011* ICSU indentifies four goals that characterise the universality of science: the universality principle; equitable access to quality data and information for research, education and informed decision-making; reaching out to all countries; and improving scientific capacity. A universal public domain for scientific data and information will be transformative for both science and society.

2.2. Data, information and ICSU

2.2.1. Shared focus

ICSU is not alone in identifying the significance of data and information to science and society. In 2010 a report of the High Level Expert Group on Scientific Data of the European Union[2] proposed a vision of:

> *a scientific e-infrastructure that supports seamless access, use, re-use, and trust of data ... [in which] ... the physical and technical infrastructure becomes invisible and the data become the infrastructure.*

In 2008 the Alliance of German Science Organisations examined a priority initiative on digital information[3] and reported that:

> *Equipping scientists and scholars with the information infrastructure best suited to meeting their research needs is the guiding principle of this priority initiative. In the digital age, this entails digital access to publications, primary research data, and virtual research and communication environments, available to the user without costs or other barriers.*

ICSU has been responsible for stimulating and coordinating scientific data and information since the 1950s when the World Data Centres were established as part of the International Geophysical Year of 1957-58. In recent years ICSU and its partners have engaged in a number of activities to improve universal and equitable access to data and information and these are summarised below.

2.2.2. Priority Area Assessment

In 2004 ICSU asked an international and inter-disciplinary group[4] to examine the broad areas in which ICSU could exercise influence in scientific data and information. This Priority Area Assessment (PAA) Panel on Scientific Data and Information recommended that ICSU should develop a strategic framework for its own activities in scientific data and information, should promote a professional approach to data and information management across all of science, and should always recall the question of who pays for data and information.

[1] ICSU Strategic Plan 2006-2011, ICSU, Paris, 64pp

[2] Riding the wave. How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. A submission to the European Union, European Commission, Brussels, 2010, 36pp

[3] Priority Initiative "Digital Information" by the Alliance of German Science Organisations, Berlin 11 June 2008. The Alliance consists of 10 German organisations.

[4] Report of the CSPR Priority Area Assessment (PAA) Panel on Scientific Data and Information, International Council for Science, Paris, 2004, ISBN 0-930357-60-4, 42pp. http://www.icsu.org/publications/reports-and-reviews/priority-area-assessment-on-scientific-data-and-information-2004/

2.2.3. Strategic Committee on Information and Data

ICSU established in 2007-08 a second group[5] with a narrower set of terms of reference, namely to guide and oversee the reform of the World Data Centres (WDCs) and the Federation of Astronomical and Geophysical Data Analysis Services (FAGS), to liaise with CODATA in the development of its strategic plan, and to advise the ICSU Committee on Scientific Planning and Review on other issues that would improve scientific data and information access and management. This Strategic Committee on Information and Data (SCID) recommended the establishment of a new ICSU World Data System (WDS) built on the foundations of the existing WDCs and FAGS, a more active and strategic role for CODATA, the development of a forum for long term discussions on data and information leading to action, and the promotion of data and information committees (where these do not exist) in ICSU National and Union Members.

Box 1. Definitions of data and information

Data and information (DI) 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 behavioral 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 them 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.

2.2.4. The Committee on Data for Science and Technology (CODATA)

CODATA was established as an ICSU interdisciplinary body in 1996. In response to the PAA on data and information CODATA developed a strategy to focus on the three key issues of the global information commons, the digital divide, and advanced data analysis methods and information technologies. CODATA has played a lead role in implementing the data sharing principles of the Global Earth Observation System of Systems (GEOSS), developing an information commons for polar science data and ensuring attention to scientific data and information issues in the World Summit on the Information Society (WSIS) and subsequent activities. CODATA is also an official member of the Consultative Committee on Units of the International Conference on Weights and Measures, addressing the role of fundamental physical and chemical constants in the International System of Units.

2.2.5. International Network for the Availability of Scientific Publications (INASP)

INASP was established by ICSU in 1992 and was registered as a UK charity in 2004. Based in Oxford and governed by an international Board of Trustees, INASP is run with a small number of full-time staff working with, and through, partners and networks in over one hundred countries. INASP supports global research communication through innovation, networking and capacity strengthening, focusing on the needs of developing and emerging countries. INASP works with partners to address their national priorities for: access to national and international scholarly information and knowledge; the use, creation, management and uptake of scholarly information and knowledge via appropriate Information and Communication Technologies; and national, regional and international cooperation, networking and knowledge exchange.

2.2.6. International Council for Scientific and Technical Information (ICSTI)

ICSTI was established in 1984 as successor of the ICSU-AB (Abstracting Board). It is an international, not-for-profit membership organisation, and a scientific associate of ICSU. Its agenda and priorities are determined by the needs and demands of its diverse membership which comprises academic libraries, scientific unions, research centres, primary and secondary publishers, learned societies, governmental organisations, funding bodies, and software and search engine companies. ICSTI fosters cooperation among all stakeholders engaged in the scientific communication process. In recent years ICSTI has concentrated its efforts on issues surrounding the 'born digital' nature of scientific communication and related technology developments. In its project-based activities, ICSTI aims in particular to initiate projects which involve the new or next-generation information users.

2.2.7. World Data System

After the acceptance of the SCID report by the ICSU General Assembly in 2008 the World Data System was established. Approximately 100 data centres and services have expressed interest in being part of the WDS. The WDS Scientific Committee has developed an implementation plan that includes a constitution, a data policy, a system architecture and a set of criteria for membership. The main action of inviting and then reviewing

[5] Ad hoc Strategic Committee on Data and Information, Final Report to the ICSU Committee on Scientific Planning and Review, ICSU, Paris, 36pp. http://www.icsu.org/publications/reports-and-reviews/scid-report/

applications for membership of the WDS was launched at the beginning of 2011. ICSU invited proposals for an International Programme Office (IPO) for the World Data System. After review and evaluation the IPO has been located at the National Institute of Communications Technologies in Tokyo, Japan.

2.2.8. Strategic Coordinating Committee on Information and Data

One of the five recommendations of the SCID report was for ICSU to establish a fixed-term group to provide broad expertise and advice to ICSU on scientific data and information management. While the detailed work of establishing a functioning World Data System is taking place, it is also beneficial for ICSU to maintain strategic momentum on the broad questions of scientific data and information management. ICSU has established a new Strategic Coordinating Committee on Information and Data (SCCID) for a period of three years. The terms of reference of the SCCID are as follows.

1. to establish and assert a visible and effective strategic leadership role, on behalf of the global scientific community, in relation to the policies, management and stewardship of scientific data and information;
2. to provide broad expertise and advice to ICSU and to ensure proper coordination among ICSU activities in the field of scientific data and information;
3. to advise on the data needs and possible solutions for existing and new ICSU programmes and other international initiatives;
4. to develop a coordinated strategy for training and capacity enhancement in data and information stewardship with a particular focus on least developed countries and involving the activities of CODATA, ICSU WDS and other relevant Interdisciplinary Bodies;
5. to provide strategic advice, where appropriate, to guide the implementation of the new ICSU WDS structure and the continued development of CODATA;
6. to work with relevant ICSU bodies and key partners to promote international discussions on current and evolving key data and information issues including global access;
7. to develop a sustainability plan for maintaining the established strategic coordination and leadership role of ICSU for consideration by the General Assembly in 2011.

The members of SCCID are listed in Appendix A. This report is the interim report of the SCCID, produced after four meetings held in Paris at ICSU HQ. The SCCID, the WDS, CODATA, INASP and the other stakeholders interact in the way shown in Figure 1 below.

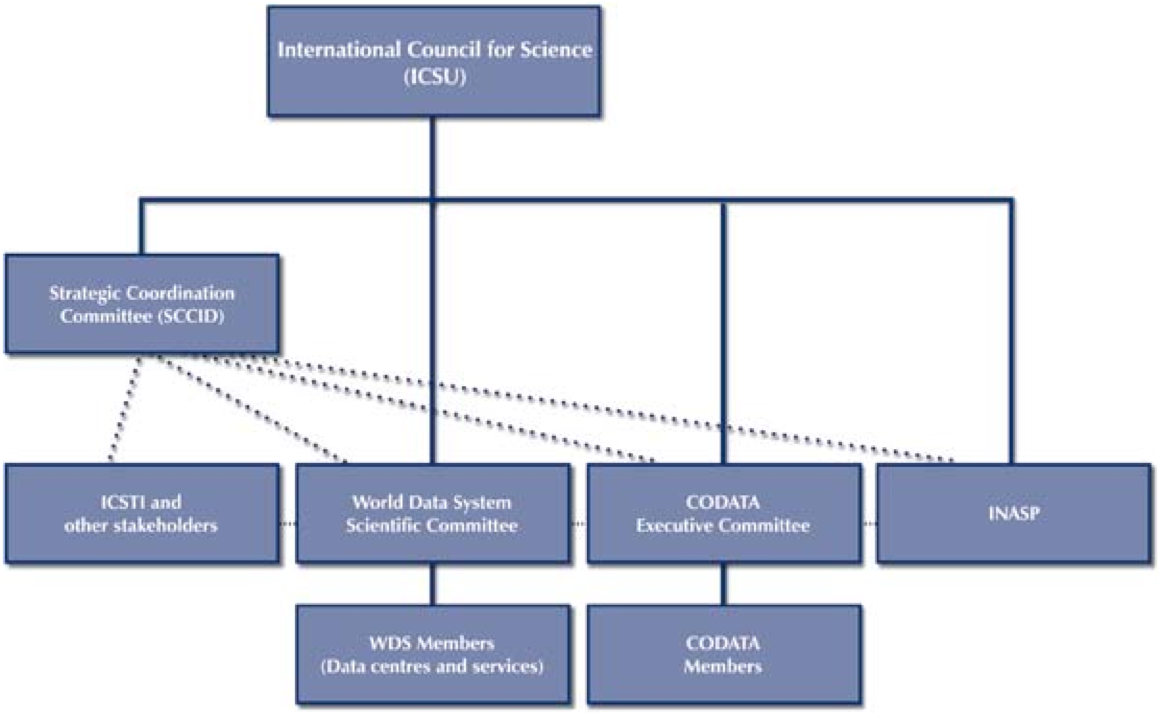
International Council for Science (ICSU)
Strategic Coordination Committee (SCCID)
ICSTI and other stakeholders
World Data System Scientific Committee
CODATA Executive Committee
INASP
WDS Members (Data centres and services)
CODATA Members

Figure 1: Interaction and relationships of SCCID, the WDS, CODATA, INASP and other stakeholders

Solid lines indicate structural links and/or direct reporting responsibilities. Lighter lines indicate strategic coordination and cooperation functions. As Interdisciplinary Bodies CODATA, ICSU World Data System and INASP have direct reporting lines to ICSU, in addition to strong cooperative links with the Strategic Coordinating Committee.

3. Data and information management challenges for science

3.1. The information explosion

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 to high energy physics. By the end of 2011 it is estimated that 30,000 human genome sequences will have been completed [6], creating information about billions of bases and requiring petabytes of data storage. A study by the International Data Corporation (IDC) in 2010[7] estimated that by the year 2020 there will be 35 zetabytes[8] (ZB) of digital data created per annum, which is 44 times the amount of digital data produced in the year 2009. The IDC estimate of the total digital storage 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 produce more digital data than is capable of being processed in all the world's computers put together. Many data sets that exist only in printed form, including those located in LEDCs, still need to be converted into digital form so that they become more widely and easily accessible.

3.2. Data complexity

There are four important characteristics of complex data; high dimensionality, multimodality, multiscale and heterogeneity. Multimode data appears in fields ranging from neuroscience to astronomy and while its origins are in imagery (for example images captured using different modalities or operating modes, channels and frequencies), 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 bio-medicine, 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 scale in animals, each of which has different measurement and structural data representations, severely inhibits system level views.

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, are beginning to provide progress in some fields. In the present context, this is all dependent on data and information and the application of data science.

Interdisciplinary science further pushes the need to accommodate substantial levels of data heterogeneity as different fields capture and represent data in significantly distinct data structures, metadata and vocabularies. For example, morphogenesis modelling can be done in a multiscale approach that spans organ, tissue, cellular and molecular levels. Semantics can play a significant role to merge the information from different scales and vertically integrate the scales. More complex analysis of multi-variable data collections results in multi-dimensional data.

3.3. 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 in some instances.

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 — a rare, valuable, yet unrecognized breed — 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[9] 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

[6] Nature 467, 28 October 2010, doi: 10.1038/nature09534

[7] IDC Digital Universe Study, sponsored by EMC, May 2010, http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm

[8] 1 Zetabyte = 10^{21} bytes

[9] NSF Data Management Plan Requirements http://www.nsf.gov/bfa/dias/policy/dmp.jsp

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[10] scandal.

3.4. Digital divide

While there are advances in data capture technologies and the ability to handle the data explosion, there is still a digital divide with those scientists in the less economically developed countries (LEDCs) lacking access to both data and technology. The World Summit on the Information Society meetings in both 2003 and 2005 identified the digital divide as a major concern for society. CODATA has identified the digital divide as a key strategic issue for attention and has a task group on 'Preservation of and Access to Scientific and Technical Data in Developing Countries'.

Figure 2 shows one version of the digital divide. It is a map of the number of computers per 100 people and was created by the UN as one of the Global Development Goals Indicators 2008. The distribution shows 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[11] shows 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 or only recently emerging from 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.

LEDCs do operationally collect and distribute data valuable to science in meteorology, oceanography and in the collection of national statistics such as population and economic data. ODINAFRICA is an example of a capacity building project to develop a cooperation network for managing and exchanging oceanographic data and information.

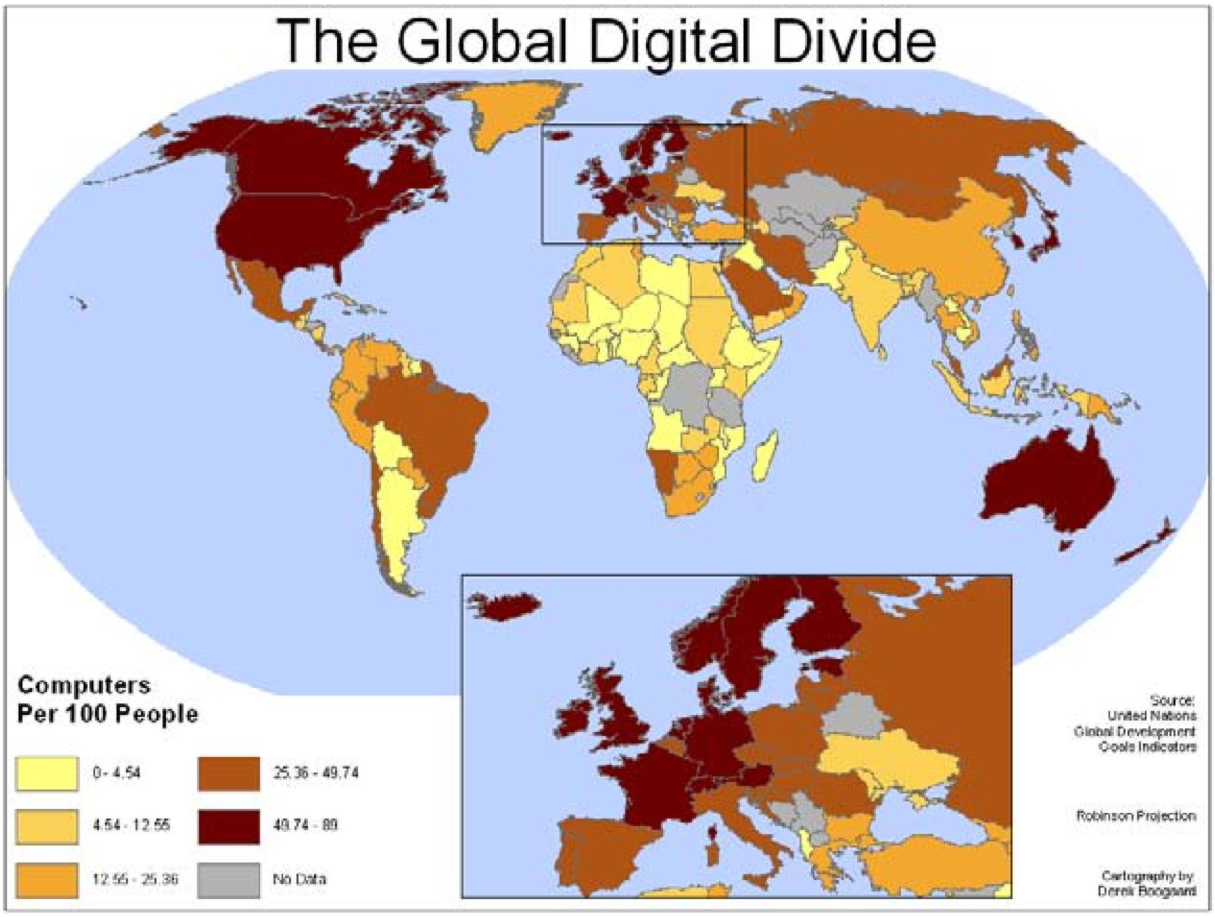

Figure 2: Computers per 100 people by country.
Source: United Nations, Global Development Goals Indicators 2008.

[10] Closing the Climategate, Nature Volume:468, Page:345, doi:10.1038/468345a

[11] ITU The World in 2009, ICT facts and figures, http://www.itu.int/ITU-D/ict/material/Telecom09_ yer.pdf

3.5. Information 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, but 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 need to provide tools to help scientists use data. Data access portals such as the IODE Ocean Data Portal and the International Virtual Observatory Alliance[12] portal simplify data access routes to some extent.

Box 2. DataCite

The consortium DataCite[12], created in December 2009 by several libraries and information centres, aims to facilitate online access to research data, increase their acceptance as legitimate, citable contributions to the scientific record and promote data sharing. A main objective is to enable scientists to locate, identify and cite datasets with confidence.

DataCite operates in particular as 'DOI registration agency', considering DOIs (Digital Object Identifiers) as a most efficient way of bi-directional linking between publications and underlying datasets. To achieve this, DataCite and its currently 15 worldwide member organisations cooperate with data centres, research institutions and scholarly publishers. Common methods and best practices are established, as for example a comprehensive common metadata scheme, and services and advice are provided, centrally as well as locally by the members.

DataCite is involved in the new CODATA Task Group "Data Citation Standards and Practices"[13].

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[13], [14]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. 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 or UniProt or to 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 (DOAJ). Traditional publishers are experimenting with new business models and increasingly offering open access options. In its November 2007 Brussels Declaration[15], the International Association of Scientific, Technical and Medical Publishers (STM) stated that "Raw research data should be made freely available to all researchers." The publisher Elsevier and the PANGAEA data centre have an agreement about connecting research articles to data sets, a service enrichment considered as "a blueprint of how Elsevier would like to work with data set repositories all over the world".

These changes are happening against the backdrop of rapid change in information technology and computer networks. Real time configurable data systems, or sensor webs, are expanding the approaches many scientists can take to experiments. Cloud computing, a major growth area of the last two years, is changing the economics of large data handling and emphasising the service of data and information provision rather than the technology itself. Intelligent software tools for data and model visualisation, data navigation and data analysis are growing at a fast rate.

It is arguable that change has always characterised science, not least because scientists create change through their results. The pace of change in data and information management is now very rapid and it is both desirable and necessary to identify and exploit the ways in which science can benefit from change so that better science is produced.

[12] See http://www.ivoa.net
[13] See http://www.datacite.org
[14] See http://www.codata.org/taskgroups/TGdatacitation
[15] Brussels Declaration on STM Publishing, http://www.stm-assoc.org/public_affairs_brussels_declaration.php

4. Solutions and strategies

4.1. ICSU leadership

The first recommendation of the SCID report in 2008 was for ICSU to assert a much-needed strategic leadership role, on behalf of the global scientific community, in relation to the policies, management and stewardship of scientific data and information. This recommendation finds an echo in the terms of reference of the new SCCID. This chapter of the report rises to the challenges set out in the earlier chapters by providing recommendations for action by the ICSU family. In this way the report delivers a way for ICSU to lead through action and evidence and not by proclamation alone.

ICSU leadership must be sustainable and so must have the capacity to endure. ICSU's engagement through the Belmont Forum with funding agencies can be fruitfully used to identify where funders can exercise influence over science data and information management and therefore encourage sustainable practices.

4.2. Interim recommendations

4.2.1. Clarification and communication of best practice

At the ICSU General Assembly in Maputo in 2008, ICSU members unanimously agreed to the five recommendations contained in the report of the Strategic Committee on Information and Data. The fifth of these recommendations was:

> *ICSU National Members and Unions be strongly encouraged to establish committees or commissions, where these do not already exist, focussing on data and information issues.*

The reason for the recommendation was to encourage the adoption of best practice in professional data management by National Members and Union Members of ICSU. In a survey carried out in early 2010, 6 of the 119 National Members and 10 of the 30 Union Members had in existence a data and information committee, although of these only one had done so as a result of the SCID recommendation. A further reason for the SCID recommendation was to enable new ICSU programmes to call upon advice about professional data management from within the national and union members.

Box 3. Example ICSU Member activities

The Board on Research Data and Information of the US National Academy of Sciences works to improve the stewardship, policy and use of digital data and information for science and the broader society. The Board proposes initiatives on scientific data and information management that might be carried out by the US National Research Council, the U.S. government, or the broader scientific community.

In an effort to increase awareness of the importance of management of research data, the Canadian Research Data Strategy Working Group is organizing a Data Summit in Ottawa in September 2011 to communicate to senior Canadian decision makers and leaders the importance of establishing a Canadian national data management strategy, including trusted digital repositories and open data access.

The International Union of Geodesy and Geophysics (IUGG) has established a Union Commission for Data and Information to provide a sustainable structure to support and strengthen IUGG science by focussing on data science, distributed data systems, open access and all other data and information management issues relevant to IUGG.

Professional data and information management in science is commonly acknowledged as necessary at all scales of scientific endeavour from that of the individual scientist to the scale of major international cooperative scientific programmes such as the International Polar Year (IPY). The open question though is: where to start? It seems to be insufficient to request professional data management without providing the right level of guidance to achieve the goal. Therefore, 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.

SCCID has produced a deliberately short, practical guide to best practice for professional data management in science and it is included in Appendix B. The guide is an initial draft and 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.

Recommendation 1. We recommend that ICSU should ensure that National Members and Union Members adopt the guide to best practice in Appendix B, either through their own data and information committees or commissions (where these exist), or independently. ICSU should also ensure that the guide is followed by all new ICSU projects and programmes. In the long term CODATA is the natural ICSU home for best practice advice on professional data management.

4.2.2. 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: the Budapest, Bethesda and Berlin declarations.

- 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 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 (see Box 1). 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.

Box 4. Information Commons

The success of community-based, Internet-enabled resources such as Wikipedia and YouTube has inspired a number of "information commons" initiatives aimed at promoting wider access to and use of shared scientific data and information. One of the most mature efforts is the Global Biodiversity Information Facility, which provides free and open access to a wide range of biodiversity data from around the world. In the life sciences, the Sage Commons focuses on developing and sharing human disease biology models and the Microbial Commons is building on the long tradition of open international exchange of plant and animal genetic material. Many of these efforts have drawn on the Protocol for Implementing Open Access Data, developed by the Science Commons Project of Creative Commons, which provides a framework for addressing intellectual property issues and developing scientific norms for managing an information commons.

Recommendation 2. We recommend that ICSU should establish a forum for the exploration and eventual agreement in relation to science of all the terms used under the broad umbrella of Open Access. The forum should investigate and compare access policies and degrees of openness in access to data and information, including the work of ICSTI and INASP in relation to publications and CODATA in relation to scientific data. The ICSU Committee on Freedom and Responsibility in the conduct of Science (CFRS) should be asked to examine how open access to data and information relates to their work. The proposed ICSU forum can use as starting points the 3B declarations cited above and the following statements on open access.

- Panton Principles for Open Data in Science, http://pantonprinciples.org
- Open Knowledge Definition, http://www.opendefinition.org/okd
- Protocol for Implementing Open Access Data, http://sciencecommons.org/projects/publishing/open-access-data-protocol
- Open Definition Conformant Data Licences, http://www.opendefinition.org/licenses/#Data
- Open Science Data, http://en.wikipedia.org/wiki/Open_science_data

The 34 members of the Organisation for Economic Co-operation and Development (OECD) have agreed at ministerial level a statement on *OECD Guidelines and Principles for Access to Research Data from Public Funding*[16]. 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. For convenience the 13 principles are included in this report in summary form as Appendix C. 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. Of the 34 OECD member states all but one, namely Iceland, is a national member of ICSU, which means that these 33 National Members of ICSU are already covered by the OECD principles for access to research data from public funding.

Recommendation 3. We recommend that ICSU uses the OECD guidelines that have already been agreed implicitly by 33 of its National Members as the basis for a forum to discuss and agree a set of principles among all ICSU National Members. It is essential to recognise that science is highly variable in its approach to access and that different models of access are appropriate to different parts of science. The earlier recommendation on an ICSU forum to explore and agree the terms used in and around open access is directly relevant here. This subject could form an agenda item for SCCID in the future.

4.2.3. Data publication

In the days when paper copies were the norm the concept of publishing one's data was largely indivisible from that of publishing the analysis and interpretation of these data. The same scientist collected, managed, preserved and analyzed the data. The formats used for data publication were quite natural and endure to this day, in the form of drawings, tables, maps and two-dimensional diagrams. Data incorporated in such publications were thereby saved for future generations, and curated in well-established libraries with long-term sustainability.

With the advent of digital data — particularly very large data sets collected by robotic instruments such as satellites, accelerators and sequencers — so-called "raw" data became accessed by an ever shrinking fraction of the users, experts that were trusted to extract the "useful" distillate of information buried in the observations. This general issue is multiplied many times if one includes the output of numerical simulations of natural systems as part of the data holdings. As a result, the gap between data curation and data interpretation activities has become wider with time. With the dominance of print as a medium for information exchange, intellectual credit became predominantly granted to the latter, and the former grew rather less visible with time.

Box 5. Legal deposit

Legal deposit is a requirement that a publisher (individual or organization) submit one or more copies of their publications to a designated repository, usually a library, within a set period of time following publication. The requirement has historically been limited to books and periodicals but is increasingly being extended, by inference or by explicit legislation, to include digital publications (which can include data and any online information). Legal deposit is a common feature in most countries of the world. Legal deposit frameworks, in terms of legislative requirements and practical process, resources and existing practice, can potentially provide solutions to addressing long term data and digital publication preservation, archiving and access challenges.

This evolution, closely linked both to scientific progress and technological advances, calls for a fresh view of the concept of "publishing" carefully vetted 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. If and when data and information could be categorised as being published (perhaps as a formal part of the World Data System) then could data and information fall under the remit of national deposit libraries? This 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.

[16] OECD Principles and Guidelines for Access to Research Data from Public Funding, http://www.oecd.org/dataoecd/9/61/38500813.pdf

At the same time, methods for citing data have evolved rapidly — for example, the DOI concept or more generally the persistent identifier. 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. Recognition is facilitated by standards and good practices, by metrics of usage and by open access. In this context we endorse the new CODATA Task Group on Data Citation Standards and Practices.

Recommendation 4. We recommend that ICSU engage actively with publishers of all kinds, including scholarly publishers, repositories and data centres, together with the library community and with scientific researchers, to document and promote community best practice in the publication of data, and to implement sustainable processes for quality assurance, long term data curation and appropriate data citation. An accompanying shift in the professional behaviour and emphasis of the scientific community is also required, to recognize and reward explicitly the publication of peer-reviewed data as part of secure and sustainable collections.

The WDS scientific conference to be held in Kyoto in September 2011 includes a session on data publication that is being convened by WDS, ICSTI and INASP: this provides an opportunity for ICSU to start showing leadership on data publication.

4.2.4. Data scientists and professionals

Unfortunately there appears to be a lack of appreciation of the need for professional approaches to the management of data over its lifecycle. As such data scientists are not valued sufficiently by other scientists. The future roles of scientists and data professionals require acceptance by disciplinary scientists who “own” the data and funding agencies who fund the acquisition/generation of data. Partnerships among the parties are urgently needed.

Box 6. Data Science

Data science is advancing the inductive conduct of science driven by the greater volumes, complexity and heterogeneity of data being made available over the Internet. Data science combines aspects of data management, library science, computer science and physical science using supporting cyberinfrastructure and information technology. As such it is changing the way all of these disciplines work individually and collaboratively. Data science is helping scientists face new global problems of a magnitude, complexity and interdisciplinary nature whose progress is presently limited by the lack of available tools and a fully trained and agile workforce.

Formal training in the key cognitive and skill areas 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.

The conduct of science research is increasingly data driven: from data assimilation to long-term time series and beyond. It is now well established that data have an intrinsic value that outlast current science foci. Unfortunately there is only part time attention given by scientists, who often choose what funding they request to address key data issues. 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 train and give data professionals more explicit recognition. In concert we must identify appropriate norms and metrics (different from scientific publications) for their review, incentives and rewards and in doing so, identify and define a community of peers.

Recommendation 5. We recommend to CODATA that it considers as the theme for its 2012 biennial conference how data science can support the delivery of the science goals of the major ICSU Earth System Research for Global Sustainability initiative and the Planet under Pressure conference organized by ICSU’s Global Environmental Change programmes planned for March 2012 in London. The CODATA biennial conference normally occurs during the autumn months, so can follow the Planet under Pressure conference in time but be linked functionally.

Recommendation 6. 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 can be used as a starting point for course development.

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.

4.2.5. Preservation and availability of information

The sustainability of data storage means that all data and information that were once acquired, formally sampled, appraised and made available to the public will be then kept safe and available in their original form without the loss of either integrity or availability. This requirement may be regarded as the key goal of a data infrastructure system. It is not yet clear to what extent the existing and emerging means of data storage comply with the formulated requirement of sustainability.

Box 7. Data Lifecycle

In science all or most of the data is acquired and stored in digital form through sensitive sensors or instruments. It should be stressed that such hardware has a crucial role to assure the quality of the data that, in turn, may determine the quality of the science. In other words, better science could not be realized without the work of those who take care of data management software, storage, high-speed networks and other components. Scientific raw data, sent from such hardware to a data acquisition computer should be stored with a (domain-specific) standardized format together with its metadata. Proper metadata is essential not only in data discovery from published data archives and/or databases, but also in data analysis leading towards new scientific insights.

In all scientific projects and programmes data management planning should be embraced at a very early stage and should be a requirement for project approval by funding agencies. Documented plans for long term stewardship of data should always be developed by projects and programmes approved by ICSU. Appendix B of this report gives practical guidance on data management planning.

Recommendation 8. We recommend that the ICSU World Data System, once fully established, increases the visibility of data centres and their data management procedures within the scientific community with the objective of enhancing the communication between scientists and data centres so that better science is produced.

Recommendation 9. We recommend that a comprehensive analysis of the advantages and shortcomings of the different data storage models and means (for example, multi-copied concentrated vs. distributed cloud systems) should be performed. The possibility of the creation of an analogue of the legal deposit system (see Box 5) for digital data and information either on the international level or by means of coordinating national efforts should be evaluated, taking into account the specific difficulties with incentives and rewards that surround data publication. The leadership of this recommendation could be offered by an existing member of the ICSU family.

4.2.6. Necessity of standards

Adoption of a systematic standards-based approach to data management will require significant cultural changes in the ICSU family and leadership by ICSU. There is a need to understand better the range of data standards used within the ICSU family and how they interrelate. Neither ICSU nor its data organizations and committees specifies a standard framework and governance for development and definition of standards by ICSU unions and initiatives.

The flow of information and ideas amongst researchers is fundamental to the discovery and innovation process, and scientific advances are increasingly built upon the work of others, captured in digital form in databases, crossing traditional boundaries in space and time and across disciplines. An integrated approach to standards will facilitate data sharing and generally improve data interoperability. Selection of appropriate standards also facilitates data discovery, visualization, life-cycle management and data re-use.

Recommendation 10. We recommend that ICSU should exploit more fully the expertise in data standards already present in CODATA, the World Data System and in its Scientific Union Members to assist in the definition and maintenance of high level data standards appropriate to meet both disciplinary requirements and overall science interoperability standards. In this connection we support the proposal on harmonisation of data standards submitted by CODATA in conjunction with four union members[17] for a grant under the ICSU Small Grants Programme to survey existing standards in use and to promote best practice on standards across ICSU unions.

Recommendation 11. We recommend that ICSU develops a better mechanism to insert a science perspective into general standards bodies such as ISO, OGC, IEEE and the World Wide Web Consortium. Suitable expertise exists in the ICSU family but it is scattered in an uncoordinated way across scientific unions and other bodies.

4.2.7. Active ICSU action in less economically developed countries

Whilst the fundamental requirements of professional data management do not differ across countries, the context and environment in which they must be applied mean that scientists and data professionals in less economically developed countries (LEDCs)[18] face specific challenges. Issues such as network connectivity, human resources capacity (both in terms of absolute numbers of people working in this area and the professional training and experience of those individuals), resource availability and networking between groups and individuals all raise particular challenges in LEDCs. Information and communication technologies are not always a cost-effective choice for many development projects. In addition to the cost of technology needed to access the Internet (for example computers, servers, modems, telephone lines, telephone usage charges), internet-based projects often require considerable training in computer and internet use. The geographic coverage of mobile phone systems is often broader and expanding more rapidly than land line availability. Therefore information and applications should be suited to these developments.

ICSU stands for and represents the universality of science. A key part of this universality today concerns equitable access to scientific data and information and so particular attention should be paid to science and scientists in countries that may be struggling to keep up with developments in scientific data and information due to their location. This feature of the digital divide requires particular focus to ensure that the divide does not widen and thereby challenge this fundamental aspect of scientific progress.

Box 8. AuthorAID

AuthorAID is a programme run by the International Network for the Availability of Scientific Publications (INASP) aiming to support developing country researchers in publishing and communicating their work. The project has two key goals: to increase the success rate of developing-country researchers in obtaining publication; and to increase the visibility and influence of research in the developing world.

The website www.authoraid.info is the key networking hub for the project. It is based on social networking sites, provides extensive learning and training resources, and researchers can register to be mentored or to be a mentor, or simply as part of a research community. In addition, regular workshops are held in AuthorAID focus countries

There is therefore a need for action and support in the following areas.

- Support capacity building and expansion of Research and Education Networks (RENs) in less economically developed countries as a key component of the scientific data and information infrastructure.

[17] IUGG, IUGS, IUCN, IUCR

[18] Less economically developed countries are those with an annual per capita Gross National Income of $995 or less http://data.worldbank.org/about/country-classifications

- Support capacity building and provide practical support for the adoption of standards and best practices in LEDCs, as outlined in this report and using Appendix B.
- Advocate the professionalization of data management in science through the development of appropriate professional training programmes and supporting structures in LEDCs.
- Support improved access to scientific information and data in LEDCs

Box 9. Science data management and information infrastructure in Chile

Chile is in the very early stages of developing a policy, using international standards, for data and information management. The main impetus for this comes from CONICYT (the Science and Technology Research Council) and the main universities. The major hurdles, which are probably common to many nations like Chile, include a shortage of qualified professionals and infrastructure, absence of formal protocols, standards and regulations to define, standardize, store and execute the processes of data and information management. There is also a lack of awareness amongst academics and other professionals concerning the importance and benefits of data and information policies.

Recommendation 12. We recommend that ICSU be actively involved in this area through the following mechanisms.

- CODATA to maintain its focus on capacity building in LEDCs in its strategic plans, including the work of its Task Groups on ‘Data sources for sustainable development in SADC countries’ and ‘Preservation of and access to science and technology data in developing countries’. Practical support for best practice and for standards demonstration projects will be particularly helpful, acknowledging the cultural differences across LEDCs. The proposal within CODATA to work with the Chinese Academy of Sciences on data and information management training is very welcome.
- The World Data System to continue to seek active nodes, data centres and services in LEDCs. The WDS should include in its agenda: opportunities for dialogue on lessons learned in data centres and services applied to LEDCs; the encouragement of data centres and services in LEDCs to become active members of the WDS; the encouragement of LEDC scientists to engage with the WDS; and the exploration of mechanisms to improve network connectivity in LEDCs.
- ICSU National and Union Members and associated bodies (in particular INASP, ICSTI and CODATA) to develop further their partnerships, mentoring, twinning and active research programmes with LEDCs to ensure that data and information management is given careful consideration.
- The involvement by ICSU Members in LEDCs in international initiatives such as the Global Earth Observation System of Systems, the International Year of Biodiversity, the International Year of Forests and the World Summit on the Information Society will encourage LEDCs to develop national data policies, share data and information, develop information services, improve network connectivity for science and enhance professional data and information management.

4.2.8. Cooperation with commercial organizations

Science has long used commercial data for its activities, and vice versa. For example, demographic characteristics, geological map data and the human genome. Increasingly there are shared challenges and solutions being explored by both the academic and commercial communities in relation to data and information management. Due to differences in approaches, motivation and ultimate outcomes, there are significant lost opportunities on both sides. At the same time, the shared interest in and advantage of information exchange are not being exploited for the benefit of science.

The economics of commercial organizations are a significant factor in possible cooperation. The impact of contrasting business models, in the form of intellectual cost and solution sharing, has been demonstrated in a few science communities as being ‘profitable’ for all. The desire from funders of science projects and programmes to attain a sustainable future raises the question of the role for commercial entities.

There is much to be done to change the culture, perceptions, and willingness for cooperation from both sides. A viable path forward would be to provide practical opportunities for collaborations (meetings of minds and needs) and to document and share successful examples of collaborations, i.e. de-mystify the commercial interest in data and information research and development in the 21st century.

Recommendation 13. We recommend that ICSTI takes a lead in this and considers enlarging its existing dialogue with Microsoft Research to include both more commercial companies and more ICSU National and Union Members to explore how science and commerce can exploit data and information to mutual benefit.

4.2.9. In-reach: the role of data in science

Individuals and groups from many different scientific disciplines and world regions have been exploring new scientific and technical approaches for generating, managing and using scientific data, and experimenting with new institutional, policy and incentive models. Yet sharing of these experiences—and efforts to put successful approaches into wider practice—have been relatively haphazard and uncoordinated. A more systematic communication and 'in-reach' strategy is needed to help improve awareness of pressing data challenges, publicize successful solutions as they emerge and build the critical mass needed to support larger-scale institutional change (for example, changes in data stewardship, data citation or professional promotion practices that cut across disciplines and countries). A key challenge is to ensure that scientists from less economically developed countries and new generations of scientists in a wide range of disciplines become aware of and have access to the tools, data and educational resources they need to succeed in data-intensive scientific discovery.

Data issues have had growing visibility in traditional communication channels such as the science press (e.g., Nature and Science), which still provides a unique channel for reaching a high proportion of the international scientific community as well as key stakeholders such as research funding agencies, data archives, libraries, museums, research institutes, science educators and the private sector involved in science. Therefore, it is important to work with representatives of the science press to improve their understanding and awareness of these issues and the innovative work being done by the science community to address them. This should include the preparation of high profile editorials and policy briefs, suggestions for interesting stories, special issues and other features, and targeted meetings and briefings on key issues and trends, involving key scientists and stakeholders.

Recommendation 14. We recommend that the World Data System be the natural home for in-reach activities. Both the WDS and CODATA should continue to raise their own profiles within science, the science press, publishers, libraries and the private sector so that experience is shared, best practice is developed and science outcomes improved.

4.3. Sustainability plan

This report has explicitly identified roles for CODATA, the WDS, ICSTI and INASP in its recommendations. They are not repeated in this section explicitly, but better communication amongst these four groups will assist in ICSU's leadership of science data and information management. All four organisations have distinctive and complementary roles to play, but often the challenges are so large, for example over the subject of data publication, that joint working will return high rewards.

While this report does contain the explicit nomination of certain groups, it should be recognised that it is the interim report of SCCID and so contains interim recommendations that can be seen as part of a process of ICSU maintaining a strategic focus on data and information issues. Deliberate and incremental progress is necessary for ICSU to develop a clearer landscape on roles and responsibilities concerning science data and information. The next phase of work performed by SCCID should focus on practical steps that fit within the strategic framework outlined in this report, including the following.

- Creating a forum for discussion and agreement on open access, building on what is already proposed in this report.
- Coordination of ICSU Union activities on data and information, and the communication of best practice across ICSU.
- Engagement with outside standards organisations to ensure that a science perspective is included in their work.
- Exploring the exciting new developments such as legal deposit libraries and data publication that have only been touched on so far.

Uncertainty in the technology and the conceptual approaches in science data and information management is likely to increase in the future rather than diminish, and therefore ICSU requires a permanent mechanism to ensure that it maintains a strong strategic presence in this important field.

Appendices

Appendix A. SCCID Committee members

Ray Harris (Chair)[1,2,3,4]
University College London
United Kingdom
ray.harris@ucl.ac.uk

Barbara Andrews[1,2,3,4]
Centre for Biochemical Engineering and Biotechnology
Chile
bandrews@ing.uchile.cl

Malika Bel Hassen[2]
INSTM Institut National des Sciences et
Technologies de la Mer
Tunisia
belhassen.malika@instm.rnrt.tn

Martin Belcher[1,2,3,4]
International Network for the Availability of
Scientific Publications
United Kingdom
mbelcher@inasp.info

John Broome[1,2,3,4]
Natural Resources Canada
Canada
broome@nrcan.gc.ca

Robert S. Chen[1,2,3,4]
Center for International Earth Science Information
Network (CIESIN)
United States
bchen@ciesin.columbia.edu

Kim Finney[1]
Australian Antarctic Data Centre
Australia
Kim.Finney@aad.gov.au

Peter Fox[1,2,3,4]
Rensselaer Polytechnic Institute (RPI)
United States
pfox@cs.rpi.edu

Herbert Gruttemeier[1,2,3,4]
International Council for Scientific and Technical
Information (ICSTI)
France
gruttemeier@inist.fr

Masatoshi Ohishi[1,3]
National Astronomical Observatory of Japan
Japan
masatoshi.ohishi@nao.ac.jp

Johan Pauw[1,2]
South African Environmental Observation Network
South Africa
johan@saeon.ac.za

Bernd Richter[1,2,3,4]
Global Reference Systems
Germany
bernd.richter@bkg.bund.de

Tieniu Tan[1,2,4]
National Laboratory of Pattern Recognition
China
tntan@cashq.ac.cn

Sergey Tikhotsky[3,4]
Institute of Physics of the Earth
Russia
sat@ifz.ru

Jean-Bernard Minster[1,2,3,4] (ex officio WDS-SC)
Scripps Institution of Oceanography
United States
jbminster@ucsd.edu

Kathleen Cass[1,2,3,4] (ex officio CODATA)
CODATA
France
codata@dial.oleane.com

The committee met 4 times from October 2009 to March 2011.
1,2,3,4 following the names of committee members indicate the specific meetings attended.

Appendix B. Principles of best practice for data and information management

This guide is an initial draft and has been produced to assist all ICSU National and Union Members and all new ICSU projects and programmes in planning and implementing their data and information management tasks. Active attention to the principles outlined in this guide will improve science by making the data and information that scientists use more readily and reliably available in both the short term and the long term.

Each point of best practice is expressed as an active verb (e.g. appoint, use, exploit) to focus attention on the action that needs to be taken.

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/dataoecd/9/61/38500813.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, see http://ipydis.org/index.html
- Research Information Network, stewardship of digital research data – principles and guidelines, 2008, 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/

Appendix C. OECD Guidelines and Principles for Access to Research Data from Public Funding

Extracts from all 13 OECD Guidelines and Principles

A. Openness

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.

B. Flexibility

Flexibility requires taking into account the rapid and often unpredictable changes in information technologies, the characteristics of each research field and the diversity of research systems, legal systems and cultures of each member country.

C. Transparency

Information on research data and data-producing organisations, documentation on the data and specifications of conditions attached to the use of these data should be internationally available in a transparent way, ideally through the Internet.

D. Legal conformity

Data access arrangements should respect the legal rights and legitimate interests of all stakeholders in the public research enterprise.

E. Protection of intellectual property

Data access arrangements should consider the applicability of copyright or of other intellectual property laws that may be relevant to publicly funded research databases.

F. Formal responsibility

Access arrangements should promote explicit, formal institutional practices, such as the development of rules and regulations, regarding the responsibilities of the various parties involved in data-related activities. These practices should pertain to authorship, producer credits, ownership, dissemination, usage restrictions, financial arrangements, ethical rules, licensing terms, liability, and sustainable archiving.

G. Professionalism

Institutional arrangements for the management of research data should be based on the relevant professional standards and values embodied in the codes of conduct of the scientific communities involved.

H. Interoperability

Although science is becoming a highly globalised endeavour, incompatibility of technical and procedural standards can be the most serious barrier to multiple uses of data sets. Access arrangements, should pay due attention to the relevant international data documentation standards. Member countries and research institutions should co-operate with international organisations charged with developing new standards.

I. Quality

The value and utility of research data depends, to a large extent, on the quality of the data itself. Data managers and data collection organisations should pay particular attention to ensuring compliance with explicit quality standards.

J. Security

With regard to guaranteeing the integrity of a data set, every effort should be made to ensure the completeness of data and absence of errors. With regard to security, the data, along with relevant meta-data and descriptions, should be protected against intentional or unintentional loss, destruction, modification and unauthorised access in conformity with explicit security protocols.

K. Efficiency

One of the central goals of promoting data access and sharing is to improve the overall efficiency of publicly funded scientific research to avoid the expensive and unnecessary duplication of data collection efforts. Consideration should be given to descriptions of good practice, data selection and appraisal, cost-benefit analysis of archives and incentives to professional data management.

L. Accountability

The performance of data access arrangements should be subject to periodic evaluation by user groups, responsible institutions and research funding agencies.

M. Sustainability

Due consideration should be given to the sustainability of access to publicly funded research data as a key element of the research infrastructure. Research funding agencies and research institutions should consider the long-term preservation of data at the outset of each new project, and in particular, determine the most appropriate archival facilities for the data.

Appendix D. Example curriculum for data science

Curriculum for Data Science taught at Rensselaer Polytechnic Institute, USA

Description

Science has fully entered a new mode of operation. Data science is advancing the inductive conduct of science driven by the greater volumes, complexity and heterogeneity of data being made available over the internet. Data science combines aspects of data management, library science, computer science and physical science using supporting cyber-infrastructure and information technology. As such it is changing the way all of these disciplines do both their individual and collaborative work. Data science is helping scientists face new global problems of a magnitude, complexity and interdisciplinary nature whose progress is presently limited by the lack of available tools and a fully trained and agile workforce.

There is a lack formal training in the key cognitive and skill areas that would enable graduates to become key participants in e-Science collaborations. The need is to teach key methodologies in application areas based on real research experience and build a skill-set. At the heart of this new way of doing science, especially experimental and observational science but also increasingly computational science, is the generation of data.

The goal is to instruct future scientist how to sustainably generate/collect and use data for their research as well as for others. Participants will learn and be evaluated on the full life cycle of data and relevant methods, technologies and best practices.

Prerequisites or other requirements

- Knowledge such as that gained in a Data Base class
- Knowledge such as that gained in a Data Structures class

Learning Outcomes

Through class lectures, practical sessions, written and oral presentation assignments and projects, participants should:

- Understand and develop skill in Data Collection and Management
- Understand and know how to develop Data Models and Metadata
- Obtain a Knowledge of Data Standards
- Develop Skill in Data Science Tool Use and Evaluation
- Understand and apply the Data Life-Cycle principles
- Become proficient in Data and Information Product Generation

Course calendar

- Week 1: History of Data and Information, Data, Information, Knowledge Concepts and State-of-the-Art, Data life-cycle for Science.
- Week 2: Data and information acquisition (curation, preservation) and metadata/provenance - management
- Week 3: Data formats, metadata standards, conventions, reading and writing data and information
- Week 4: Class exercise - collecting data - individual
- Week 5: Class Presentations: present your data 1
- Week 6: Data Analysis and Visualization
- Week 7: Data Mining
- Week 8: Class Presentations: present your data 2
- Week 9: Class exercise - group project - working with someone else’s data notes
- Week 10: Class Presentations: present your data 3
- Week 11: Academic basis for Data and Information Science, Data Models, Schema, Markup Languages and Data as Service Paradigms

- Week 12: Data Workflow Management and Data Stewardship
- Week 13: Webs of Data and Data on the Web (Semantic Web, Linked Data), the Deep Web, Data Discovery, Data Integration
- Week 14: Project Presentations

Assignments

- Assignment 1: Preparing for Data Collection
- Assignment 2: Presenting your Data
- Assignment 3: Reformatting Data for Preservation
- Assignment 4: Working with someone else's data (team project)
- Assignment - final: Stewardship: Workflow construction for Data Life Cycle

ICSU
International Council for Science

Strengthening international science for the benefit of society

5, rue Auguste Vacquerie
75116 Paris, France

Tel: +33 1 45 25 03 29
Fax: +33 1 42 88 94 31
secretariat@icsu.org

www.icsu.org