Background Information and Terms of Reference for the ICSU Priority Area Assessment on Scientific Data and Information

Introduction

The goal of the Priority Area Assessment (PAA) process is to strengthen ICSU's overall capability in addressing priority scientific issues that are of emerging importance to science and the society at large. The PAA is a mechanism to develop ICSU's strategies for selected priority scientific areas. It is designed to help ICSU to develop a programme structure reflecting its priorities; to ensure synergies in the activities of the ICSU family; and to enable an appropriate allocation of limited resources. In order to be effective, the PAA process must involve relevant members of the ICSU family – *i.e.* interdisciplinary bodies, joint initiatives, and Union and National Members. It should consider ICSU's priorities in the context of relevant activities outside ICSU.

The immediate outcome of a PAA is a report and key recommendations that will be published and widely disseminated by ICSU. This report will form the basis for future actions by ICSU and ICSU family members, including the development of new research projects or programmes, the development of policy initiatives and definition of new priorities for the ICSU grants programme. Some of the recommendations may require the establishment of new partnerships with bodies outside the immediate ICSU community or may be more appropriately taken forward by other organizations, in which case the necessary dialogue(s) will be initiated.

The 27th General Assembly (**27th GA**) of ICSU, which was held in September 2002, endorsed the recommendation of the Committee on Scientific Planning and Review (**CSPR**) to carry out a Priority Area Assessment on Data and Information[1]. A scientific forum on this topic was held as part of the 27th GA and a summary report of this forum is attached at **annex 1**.

Context for the PAA on Scientific [2]Data and [3]Information

[1] Other PAAs are also being initiated in the areas of "capacity building" and "the environment in relation to sustainable development".

[2] Scientific data as defined here is 'raw' relatively unprocessed data from scientific experiments or observations. This is usually ordered in the form of databases and may be variously standardised or evaluated, e.g. statistically. A particular focus of the current review is large data-sets, including genomic databases, demographic and geophysical data records.

[3] Scientific information is processed and analysed data, for example in the form of scientific publications.

It is recognized that there is a continuum between data and information and for the purposes of this review strict definitions are perhaps not necessary.

All scientific disciplines are dependent on the production and use of data and the integration of data and information from different disciplines and across national boundaries is central to ICSU's role in supporting international science for the benefit of society. New information technologies offer tremendous opportunities for the improved communication, exploitation, storage and processing of scientific data and information. At the same time new challenges are appearing, relating to the commercial value, accessibility, security, standardisation and validation of scientific data. New mechanisms and models are being developed for the publication and dissemination of scientific information. The field is moving incredibly quickly due to both the 'pull' of new technological developments and the 'push' of science needs. Despite its rapid evolution, technology itself is still a limiting factor in many areas of science and major paradigm shifts are appearing, *e.g.* the development of the distributed GRID as opposed to the building of even bigger super-computers. Moreover, the necessary institutional structures and legislative frameworks are not always in place to ensure that the optimum benefit can be derived from scientific data and information for the benefit of science and society as a whole. These problems are exacerbated in the least developed countries (or in certain regions of the so-called developed countries) where infrastructure and necessary human resources are lacking. The phrase 'digital divide' has been coined to describe this situation at a societal level but there are also expanding 'digital divides' within the international science community. The **Universality of Science**, including access to data and information for all scientists is central to ICSU's mission.

This PAA on data and information is particularly timely because the World Summit on the Information Society (**WSIS**) will take place in 2003 and 2005. Many of the challenges to science in the new information society cannot be resolved by scientists alone and it is therefore crucial that the scientific community engages in a dialogue with governments the private sector and other sectors of civil society. What is the role and position of scientists in the new information society? Whilst the WSIS is not the specific focus of this assessment, the occurrence of this Summit may provide a unique opportunity for scientists to influence the development of the information society for the benefit of science and society.

Scope of the PAA on Scientific Data and Information

ICSU has to focus its own efforts, where it can have most impact. In this regard, key criteria for ICSU are 'international' and 'interdisciplinary' and these concepts are important in defining the focus of this review at the primary level. Data and information issues that are limited to single scientific disciplines and countries would generally not require the involvement of ICSU (although exchange of experiences and 'best practice' between disciplines are certainly of relevance). At a secondary level one might also consider two generic areas where ICSU currently plays a role: 1) policy development and 2) management/monitoring of key policy areas.
It is likely and entirely appropriate in a broad forward-looking strategic Assessment that priorities will be identified where the primary actors are other than ICSU. In this case, these will be communicated to the appropriate bodies or the necessary partnerships with other international organizations can be developed.
It will be important that the review committee itself, as one of its first tasks, considers the scope of its work and defines the boundaries of the review; for example Intellectual Property Rights (IPR) and copyright issues relating to digital data are increasingly important but are currently not dealt with by ICSU.

Major current ICSU Activities in relation to Data and Information

Data and information issues pervade all scientific endeavours and, as such, are of concern to the entire ICSU family – International Science Unions, National Academies, interdisciplinary bodies, joint initiatives and associates.

ICSU has one Advisory Committee – the Committee on the Dissemination of Scientific Information (**CDSI**) whose special focus is scientific publishing and access to the scientific literature- both at a policy development and management level. The international Network for the Availability of Scientific Publications (**INASP**) is a major programme of CDSI and many of the ICSU Unions also publish their own scientific journals. The International Council for Scientific and Technical Information (**ICSTI**) and the International Federation for Information Processing (**IFIP**) are very active associate members of ICSU.

ICSU has several interdisciplinary bodies, whose principal focus is the management and use of large scientific data sets: the Committee on Data for Science and Technology (**CODATA**), the Panel of the World Data Centres (**WDC**) and the Federation of Astronomical and Geophysical Data Analysis Services (**FAGS**). The Scientific Committee on frequency Allocations for Radio Astronomy and Space Science (**IUCAF**) has a very focussed co-ordination role relating to the activities of the WDC and FAGS and the Global Observing Systems (GOS)[4], which are co-sponsored by ICSU. Either individually or working together, many members of ICSU are very active with regards to scientific data. An important initiative has been the creation of the Inter-Union Bioinformatics Group **(IUBG)** which recently produced an authoritative report and recommendations regarding biological databases.

In summary, the ICSU family has many single discipline or topic focussed activities in which scientific data and information use are a major component. It also has several overarching committees with special responsibility for data and information. The future role and function of these latter committees in, in particular, should be a key aspect of this review.

The Assessment Process

The ICSU Committee on Scientific Planning and Review (CSPR) is responsible for all Priority Area Assessments and will appoint an *ad hoc* PAA Panel for each major priority scientific area (in this case 'scientific data and information').

Membership of Panels (10-15 members in total) will normally include:

- International experts with wide knowledge of the area, including some individuals with knowledge of the relevant ICSU interdisciplinary bodies, joint initiatives and ICSU Member activities. These experts to be proposed by CSPR after consultation with all ICSU member organizations;
- One CSPR member; and
- One or two additional member(s) suggested by the ICSU Executive Board (**EB**), at its discretion.

Whilst appropriate expertise must be the primary criteria for appointment to a panel, due attention will be paid to geographic distribution, disciplinary expertise, age and gender in the final panel composition.

The **terms of reference** and **proposed membership** of the panel for the PAA on Scientific Data and Information are given at annexes 2 and 3.

[4] The assessment of WDC, FAGS and the GOS should be limited to data management and policy issues. Their contribution to environmental research will be considered in the PAA on "Environment and its Relationship to Sustainable Development".

Annex 1

[5]Report from the Scientific Forum on Insuring Global Access to Scientific Data and Information, ICSU General Assembly, Rio de Janeiro, September 2002

New information technology offers tremendous possibilities for the communication, exploitation, storage and processing of scientific data and information. At the same time, new challenges continue to appear. Data produced through research often has economic value in modern society. Some research is financed by the private sector, which complicates the access, dissemination and preservation of these data. In other cases, governments wish data collected under public funding to be commercialized. There is no consensus on how international and global monitoring systems should be financed. The increasing value of biological data is apparent to all. New information technology has motivated the creation of new intellectual property rights (IPR), such as the *sui generis* right in the European Union. The distinction between copyright and *sui generis* right does not, however, match the distinction between information and data.

Topic 1: Research databases

Background: *Data and Scientific Discovery.* Data preservation has been a major preoccupation of human society both to save collected data and to support scientific discovery. Several early civilisations used architecture (*e.g.* pyramids) for archiving astronomical information. Until recently, the simplicity and limited size of data collections allowed them to be fully accessible to individual researchers. Today, the complexity and very large size (terabytes) of databases has fundamentally changed the interaction of an individual with modern data collections. In fact it is suggested that future scientific discovery will be critically dependent on computers, and that the search for scientific discoveries may be computer-supported or even automated. Access to accurate and complete data collections is therefore important. This fundamental change in discovery is a major driving force in making data access an important issue for ICSU to address.

Current problems*: New IPR legislation for databases.* Data acquisition is a complex and expensive process, and there is no consensus on the conditions for making data available to the scientific community. IPR issues make it difficult to exchange data. IPR for databases in the European Union (EU) area is subject of a directive that was enacted in 1998. This directive defines a new *sui generis* right preventing extraction and/or reutilization of the whole or substantial parts of a database for a period of 15 years, renewable. Exceptions may allow database extraction for illustration, teaching or scientific research, but these exceptions are optional and are decided by each member state. Possible directive changes can make even more complex the accessibility of data and information for scientists and may impede normal research activities.

CDSI and CODATA activities*: Interaction with EU authorities*. The ICSU/CODATA *ad hoc* group on data and information analysed the effects of this directive, making comparison with other legislation. The group observed that there is a real need for a short statement explaining what the problem is and what could be the effects of new directives in this matter. Subsequently, there have been several contacts with EU authorities in order to explain the position of ICSU, as expressed in the resolution on ***'Principles for Dissemination of and Access to Scientific Data'***.

[5] The agenda for this Forum is at appendix 1. This report was produced by Erik Sandewall, Jean-Jacques Royer and Roger Elliott, for reporting back to the plenary session of the ICSU General Assembly. This formed the basis for a resolution at the General Assembly, which is at appendix 2.

Scientific Data Issues in the Biological Sciences. The Inter-Union Bioinformatics Group (IUBG) pointed out in their report that new discoveries in biology, especially in the genomics area, clearly demonstrate the need for a primary data repository in public archival database, even for commercial data, in order to safeguard them for the future. This requires long-term commitments. Additional work is needed on standardization issues with respect to vocabulary, hierarchy of terms, gene ontology, format and data exchange.

Topic 2: Research Knowledge Management

Background: The use of computer-based document preparation, whereby authors have taken over a large part of the activity previously done by typesetters, as well as communication by electronic mail and the easy dissemination of research articles using the world-wide web, has radically changed all steps in the publication chain for research articles. This is often perceived both as an opportunity for better and less costly communication in the research community, and as a threat to organizations that are traditionally involved in the publication chain. At the same time, it is clear that only a limited part of the possibilities of new technology are presently being used in this context.

In particular, a large number of free-access research journals have been started in recent years, allowing readers to access articles from web pages without subscription or other charges. In some cases, these new journals have also pioneered new ways of operating the peer-review process and introduced author-reader interaction. It has been presumed that the cost of production of the journal will be orders of magnitude smaller under this system, but the question of how to finance the remaining costs has not been resolved in a definite way.

Perspective: This topic was addressed by adopting the perspective of how the scientific community as a whole manages its needs for internal communication and for knowledge management (borrowing a term that is widely used in industry). The present publication system can be understood as though our community 'out-sources' the solution to its internal communication needs, and at high cost. The emergence of free-access journals can then be interpreted as a way of 'in-sourcing' that same function, and we need to discuss whether which approach is the most efficient from a total system point of view.

The data and information of concern here appears in three steps: the raw data that precede the publication of articles; the research articles themselves; and the information that enables readers to select articles that are relevant for them. The great importance of the third type of information was emphasized, and a number of ways of extending it using modern information technology were proposed.

With respect to how the costs in the publication chain could best be covered in the context of the new technology, it was proposed:

- that the preparation costs for the documents, including both typography and language correction, could best be met by the author's home institution, often on a do-it-yourself basis;
- that the costs for posting articles and making them publicly available over the Internet ought to be considered as infrastructure, to be covered on a common basis, in a way similar to how university libraries are usually financed;
- that the cost of the selection-supporting information, ranging from classical peer-review to discussion groups, dynamically updated bibliographies, etc. could often be covered in ways similar to computer freeware, that is, by a combination of voluntary efforts and direct support by parties that have a concrete interest in the proper functioning of these activities. This would

include, in particular, the research funding agencies, but other possibilities were also suggested.

It was argued that a freeware approach to the third type of information would be the most conducive to the emergence of new methods and new services.

It emphasized, finally, that such an organization of research publication could only be seen as a complement to the traditional one, and that for the foreseeable future it should co-exist with the conventional, subscription-based publication system.

Topic 3: Worldwide dissemination of scientific information

Background*: Developing Countries' Problems in Accessing Data and Information.* The digital divide is increasing, extending the knowledge gap and making it even more difficult for developing country scientists to access data and information. Usually, developing countries have limited and low quality access to the digital community, due to a lack of financial resources and of infrastructure. The Industrially Developed Countries invest on the average 2.5% of their Gross Domestic Product (GDP) for research, compared to less than 0.6% for Third World Countries (TWC). The conditions are very different between developed and developing countries: access to telephones differs by a factor of 4, network connection by a factor of 200, and Internet usage by a factor of 2000. There is a need for skilled manpower, for low-cost information access, for journals and for contacts. Recent initiatives by publishers offering favorable pricing rates for TWC must be encouraged. Some recommendations are to encourage TWC scientists to publish in electronic form, to exploit existing electronic archives, *e.g.* InterAcademy Panel for International Affairs (**IAP**), Third World Academy of Sciences (**TWAS**), International Society for Computers and their Applications (**ISCA**), and to persuade publishers to give them free or strongly discounted electronic access to scientific journals.

CDSI and CODATA activities. CDSI initiated in 1992 an International Network for the Availability of Scientific Publication (**INASP**) to promote the diffusion of scientific information in an accessible form in developing countries through newsletters, handbooks, manuals, web sites on general science, health and rural development, library support, etc. INASP aims to deliver information, to raise awareness, to disseminate national and regional information, and to strengthen local scholarly publishing. Its most recent initiative, the Programme for Enhancement of Research Information (**PERI**) provides core journals from international publishers and considerably reduced rates.

Issues for the future

Many questions still remain with respect to scientific data and information. To what extent will commercial consideration complicate access, preservation and control of information? ICSU-CODATA have made recommendations concerning the importance for the scientific community of having the right of free access included in the EU database directive. Scientific endeavor on the international level needs to be able to use the exception clause in that directive. Attempts to energize national members on this issue met were not very successful. IUBG analyzed medical genome data; CODATA can encourage code development and standardization. New technology is changing the world, science can help keep technology under control and has the will to do that. But science needs to use new structures and new modes of operation. How the opportunity of new technology can amplify the access to information is a fundamental question for the future. An important challenge for ICSU is to provide a continuous funding programme that does not depend too much on a single source.

Appendix 1:

Agenda for the scientific forum at the 27th ICSU General Assembly

3. Ensuring Global Access to Scientific Data and Information

Data issues are becoming more and more important. Research is increasingly being financed by the private sector, and data access is often difficult. In many cases, governments wish data collected by public funds to be sold. At the same time, there is no consensus on how the international and global monitoring systems should be paid for. IPR issues make it difficult to exchange data, and there is a EU directive for the legal protection of databases. The digital divide is increasing, making it even more difficult for developing country scientists to access data. These issues will be addressed by ICSU in the preparation of the World Summit on the Information Society (December 2003)

Chair: Roger Elliott (CDSI/ICSU Press)
Rapporteur: Jean-Jacques Royer (CODATA)

09:00	Opening Remarks	Roger Elliott (CDSI/ICSU Press)
09:10	Collection, Archiving & Dissemination of Global Databases	John Rumble (CODATA)
09:40	The EU Directive on Legal Protection of Databases	Ferris Webster (WDC)
10:10	Standardization and Access to Biological Data	Jean Garnier (IUBG)
10:40	Coffee Break	
11:00	Should Scientific Information Be Free ?	Erik Sandewall (CDSI/ICSU Press)
11:30-	The Digital Divide: Overcoming the Knowledge Gap	N. Mukunda (CDSI/ICSU Press)
12:00	Information Access in Developing Countries	Carol Priestley (INASP)
12:30	Discussion	
12:50	Summary and Conclusions	Roger Elliott (CDSI)
13:00	Close of Session	

Appendix 2

Resolution accepted by the 27th General Assembly of ICSU, September 2002

8. Ensuring Global Access to Scientific Data and Information

The 27th General Assembly of ICSU,

Noting the tremendous impact of new Information and Communication Technology (ICT) on the acquisition, exploitation, processing, communication, storage and long-term availability and preservation of scientific data and information, calls on all ICSU members to maximize the use of ICT for the benefit of the scientific enterprise;

Emphasizing that an efficient and well functioning infrastructure for information and communication technology is of the utmost importance for science and the dissemination of data and information worldwide;

Accepts the recommendations of the IUBG Inter-Union Bioinformatics Group in relation to biological data;

Encourages all ICSU members to use the new possibilities to obtain greater control by the scientific community of the information chain, through use of alternative forms of publication of scientific articles; and

Further noting the special challenges encountered by developing countries in making optimal use of the new facilities,

Applauds the activities of the International Network for the Availability of Scientific Publications (INASP) and similar programmes which make production of, and access to, data and information more available across the digital divide;

Encourages the relevant members of ICSU to identify the issues which the scientific community would wish to bring to the World Summit on the Information Society, and charges the Executive Board to ensure the timely input of the scientific community, in cooperation with appropriate partners, to this Summit; and

Requests that the Executive Board take all appropriate action to safeguard international databases and to ensure that full and proper access is maintained.

Annex 2

Priority Area Assessment for Scientific Data and Information

Terms of Reference

1) Define an overarching "mission" and role for ICSU in the area *of "Scientific Data and Information"*, taking into account relevant activities outside of ICSU;
2) Propose a strategic framework for ICSU to take this area forward for the next 5-10 years;
3) Examine current activities within the ICSU family; identify gaps, overlaps and synergies of existing activities, and; propose responsibilities for individual bodies;
4) Propose modalities for promoting collaboration and co-ordination within the ICSU family when necessary and propose potential partnerships with bodies outside ICSU;
5) Examine and propose, if appropriate, changes either in the future direction of individual bodies or the way they operate including relationships with other bodies/organizations.

Two additional ToR for this specific review on" data and information" are also proposed as follows:

6) To consider the ethical issues related to scientific data and information and, where necessary, propose how ICSU might develop policies in response to these issues.
7) To identify policy issues of particular importance to science and society, which should be highlighted in the World Summit on the Information Society.

Work plan

A minimum of two physical (2-day) meetings of the Panel might be expected in addition to 'virtual' discussions via e-mail and telephone. In-put from the relevant members of the ICSU family will need to be solicited and analysed; it may be desirable to conduct 'face to face' interviews with representatives of key bodies. The first meeting might focus on future developments and priorities in the field as a whole. The second meeting could be more focussed on existing and future ICSU activities and structures. Additional meetings can be planned as necessary.

The Panel will be asked to prepare a report to the CSPR, which will include an overarching mission statement and strategic framework for ICSU and recommendations on roles of new and/or existing interdisciplinary bodies and joint initiatives in the area. This report will be published.

Resources

ICSU will provide financial resources to carry out the review, including travelling and accommodation costs for the Panel members to participate in the two meetings. The ICSU Secretariat will provide administrative support to the Panel, including assistance for communication among the members and organization and arrangements of meetings. The final report will be the responsibility of the panel although the Secretariat will assist in its preparation as necessary.

Annex 3

Proposed membership of the ICSU Priority Area Assessment panel on Scientific Data and Information

Roberta Balstad Miller, Chair, Columbia, USA [Demography/social sci.]

Jean Bonnin, Strasbourg, France [geophysical data]
Marc Brodsky, American Institute of Physics, USA [publishing]
Graham Cameron, EBI, Cambridge, UK [Bioinformatics]
Liu Chuang, Beijing, China [Geographic data/remote sensing]
Carlos Correa, Buenos Aires, Argentina [Law/economics/IPR]
Norihisha Doi, Keio University, Japan [computer science]
Ray Harris, University College London, UK [earth observation data]
Andrew Kaniki, National Research Foundation, S. Africa [Information sci.]
Vitaly Nechitailenko, Moscow, Russia [Geophysics/publishing]
T B Rajashekar, Bangalore, India [Information Science]
* Pierre Ritchie, Ottawa, Canada [Psychology]

In addition to the principle expertise noted in parenthesis, members have been selected for their broad knowledge and experience of scientific data and information issues.

*member of the ICSU Committee on Scientific Planning and Review (CSPR)