Response to the ICSU "Priority Area Assessment on Scientific Data and Information"

27 October 2004

Ray Norris, Brian McMahon, Krishan Lal

CONTENTS

1. Executive Summary

The ICSU PAA on Scientific Data and Information presents a vision of full and equitable access to, and the effective use of, scientific data; and challenges the community of science to work towards that vision. An important element of the challenge is to identify and facilitate the technical and policy developments that are necessary to achieve this vision, and also to identify and mitigate the obstacles to its realisation. By virtue of its long-established role as a forum for addressing interdisciplinary and international scientific data issues, CODATA is in an excellent position to offer leadership in the areas of interoperability, data quality, long-term preservation, access and management. In this report, we suggest a CODATA response to the PAA. Those in **bold** below are actions that the CODATA Executive Committee might choose to take immediately.

We recommend to the CODATA General Assembly that:

1. **CODATA should congratulate ICSU on a comprehensive and visionary review, and should signal its desire to play a key role in realising that vision (Rec. 1)**
2. **CODATA should focus its strategic directions by:**
   - **declaring its objectives clearly through a mission statement on its home page (Rec. 2),**
   - **developing a five-year Strategic Plan of action in support of its mission (Rec. 3),**
   - **establishing a technical advisory committee to help identify interdisciplinary objectives that are aligned with CODATA's strategic goals (Rec. 4),**
   - commissioning new Task Groups to address strategic goals, while retaining the existing Task Group mechanism to allow independent initiatives to be introduced, endorsed and sponsored (Rec. 5),
   - investigating new routes for funding project-based and long-term activities to enable expansion in their scope and effectiveness (Rec. 6),
   - monitoring developments in intellectual property rights, providing recommendations of best practice within science, and contributing to the ICSU framework (Rec. 7),
   - identifying specific areas where the creation and maintenance of scientific data, or access to and use of the data are hampered by a shortfall in local resources and capacity (the "Digital Divide") and promoting technical and policy initiatives to address these (Rec. 8),
   - promoting the principle of full and equitable access to scientific data and information, both in its published mission statement and in its activities (Rec. 9; PAA #36),
   - investigating the adoption of community standards for the validation and verification of scientific data and publications to ensure the highest achievable data quality (Rec. 10),
   - participating in eScience forums, and building links to eScience groups, to help build eScience on firm data management principles (Rec. 11).
3. CODATA should engage with other members of the ICSU family (Rec 12) in addressing specific recommendations of the PAA, such as:
   - seeking full and active representation on SciDIF (PAA #58),
   - helping establish cross-disciplinary metadata standards for the classification of scientific data to encourage interoperability between emerging metadata standards in different scientific disciplines (PAA #22, 32),
   - promoting policies and appropriate technical implementation of long-term access to scientific data and information (PAA #27, 13, 14),
   - exploring the role and practice of data confederations and distributed data systems, and seeking to understand how they can effectively interact with centralised data resources (PAA #33),
   - strengthening relationships with bodies such as ICSTI and INASP by exchanging reports, conducting joint workshops, inviting representatives to meetings, and encouraging active collaboration on specific projects of common interest.

2. Background

In early 2004, ICSU commissioned a panel of experts to perform a Priority Area Assessment (PAA) on Scientific Data and Information. The resulting report is strategic and visionary. It identifies a number of shortcomings in the way that the scientific community currently manages data and information, and proposes a series of measures to ensure the healthy management of data and information in the future.

A number of issues discussed in the report are central to the role and mission of CODATA. The report notes the significant contributions of CODATA in the past, and makes a number of recommendations on the ways in which CODATA may play a significant role in the future in reshaping the way that the scientific community manages data and information.

Although the PAA's comments on CODATA are generally positive, there were some misunderstandings and omissions. Some of these were addressed in CODATA's formal response to an earlier draft of the PAA report (letter of 25 May 2004 from Professor Shuichi Iwata to Roberta Balstad Miller, Chair of the PAA). Others have been identified in the response of the US National Committee (submission by Julie Esanu and Paul Uhlir) and are included within the proposals of Appendix 3 for a memorandum in reply to the PAA Committee.

CODATA must now consider how to respond to the PAA report, and how best to contribute to achieving the goals identified in the report. A small committee consisting of the authors of this report was established at the invitation of Prof. Iwata, President of CODATA, to consider the question: "*Given the mandate of CODATA and the financial and human resources currently at its disposal, what suggested actions could or should CODATA begin that would directly/indirectly contribute to the recommendations outlined in the report*." The intention is that the committee should arrive at some specific recommendations that can be put before the CODATA meeting in Berlin in November 2004.

The group first met by teleconference on 1 October 2004, and subsequently continued discussions by email and teleconference, leading to this report.

Note: The term "Open Access" has a number of different connotations, sometimes leading to ambiguity or misunderstanding. Throughout this report, we use the phrase "equitable access" to imply equitable and non-discriminatory access to data and information, while preserving commercial value where appropriate, and with an appropriate funding model that ensures accessibility in developing countries. We use the phrase "Open Access" to imply that information or services are available at zero or minimal charge.

3. A Suggested CODATA Response

3.1 The Role of CODATA

The PAA report is built around a number of strategic recommendations to ICSU and its associated organisations, and encapsulates a vision of how scientific data and information may be managed effectively in the future. This is an area in which CODATA by its nature is a major stakeholder. Many of the PAA recommendations (*e.g.* #5 – establish guidelines, and #57 – establish a strategic framework), though directed at ICSU, cover areas in which CODATA could, and in the opinion of this committee, should, offer to play a significant role. CODATA can act as a key mechanism by which ICSU implements these recommendations. We recommend that CODATA should embrace the goals portrayed in the PAA report, and respond positively and proactively to the report. This will involve significant additions to the aims and objectives of CODATA, broadening its scope from a forum for discussion of data issues to a group in pursuit of a strategic goal.

Recommendation 1. CODATA should congratulate ICSU on a comprehensive and visionary review, and should signal its desire to play a significant role in realising that vision, including playing a key role in the proposed Scientific Data and Information Forum SciDIF.

3.2 The CODATA Mission

Many of the activities of CODATA are directed at strategic goals, and there is clear support from the CODATA membership (through the endorsement of the General Assembly) for these activities and for their effective and energetic execution. Nevertheless, these goals tend to be implicit in the direction and activities of CODATA, rather than being explicitly expressed in its published aims and objectives. As a result, the strategic aims are not evident to outside observers (such as the authors of the PAA), and it is difficult to formulate strategic objectives and monitor progress towards achieving them. We believe that a mission statement will help clarify and focus these aims. It should encapsulate the values of the organisation and its broad purpose in promoting the use and management of scientific data for the common good. We also recognise that formulating a good mission statement is a major undertaking, and needs to include input from all significant stakeholders.

Recommendation 2. CODATA should distil its aims and objectives, together with any additional goals identified by its members and other stakeholders, into a concise mission statement to be published on the CODATA web home page.

3.3 The CODATA Strategic Plan

In line with the mission statement, a Strategic Plan should be formulated and published that describes the specific actions that will enable CODATA to achieve its goals or mitigate threats that might prevent their realisation. The Strategic Plan should be specific, contain achievable objectives, and concentrate on a finite period (*e.g.* five years). It will also include long-term objectives which are not achievable within that time frame, but whose progress should be measurable within the period. Regular reporting of CODATA activities within the framework of a Strategic Plan will help to give a more coherent view of directed progress in pursuit of the organisation's mission goals. It will also assist in the assessment of CODATA's performance within the broader strategic framework established by ICSU as a result of this PAA review.

In Appendix 1, we suggest a possible outline and implementation of the Strategic Plan.

Recommendation 3. CODATA should develop and publish a five-year Strategic Plan to direct its activities in pursuit of its mission objectives.

3.4 Technical Advisory Committee

Since its inception, CODATA has been an energetic multidisciplinary body. The biennial CODATA conference is one of the organisation's most visible and most successful activities, and provides an unmatched cross-disciplinary forum for bringing together expertise on data issues across all of science. However, there is currently no mechanism for identifying and implementing technical solutions to challenges that are common to different disciplines. This represents a significant opportunity for CODATA to play a leading role in promoting good data management across all sciences and to further the aims of CODATA and ICSU.

We suggest in Appendix 2 a number of examples that are potentially achievable by, or under the direction of, the resources at CODATA's disposal. These are not intended to be exhaustive or definitive, but rather to act as a starting point for further thought.

We believe it would be useful to set up a "think tank" of members who are alert to technical developments and can suggest their deployment in interdisciplinary projects of this sort. The output of this group would be suggestions to the Executive for possible Task Groups, Workshops or other mechanisms (such as web sites) to investigate projects of the sort suggested in Appendix 2. Initially we propose that this take the form of a technical advisory committee to the Executive. As such, the existence of the group would be at the discretion of the Executive, allowing it longevity to carry out its activities so long as it was working effectively, and with the possibility of winding up without formal action when it was felt to have served its purpose or was no longer effective.

Recommendation 4: CODATA should form a technical advisory committee to the Executive to identify technical developments that would further the objectives of CODATA through specific Task Groups, Workshops or other activities.

3.5 Task Groups

Project-based Task Groups have proved very successful for tackling well-defined problems, and many CODATA Task Groups have produced important results disproportionate to their size. The current mechanism for proposing Task Groups enables an initiative to be advanced by a member, subjected to peer review and Executive recommendation, with final approval by the General Assembly. Subject to the availability of suitable resources, it provides a mechanism for innovative research and cooperation in international and interdisciplinary ventures, and encourages involvement from the entire membership. Despite the PAA comment of "opportunism", we believe that the existing mechanism for appointing Task Groups has resulted in many valuable projects that are consonant with and enhance CODATA's aspirations. We therefore recommend that this process should continue.

In addition, under the umbrella of a Strategic Plan, it is likely that there will be opportunities for the Executive to commission new Working Groups or Task Groups. Historically, Working Groups have been set up by the Executive to investigate specific problems. They may or may not develop into Task Groups. Although it is not a rigid definition, often Working Groups characterise the investigative phase, and Task Groups the active phase of a project. Working Groups do not require the formal approval of the General Assembly. They therefore allow the Executive flexibility in directing action on specific issues. Since they are not accountable to the General Assembly, they are, in general, best employed as short-term measures. Task Groups are subject to the approval of the General Assembly, and are therefore accountable to the membership as a whole. We commend this practice of dual categories of task force for its flexibility.

The ICSU PAA report suggests that "Given its limited resources, some of the historical CODATA activities should be terminated or principal responsibility for them transferred elsewhere" (p. 44 of the draft PAA report). To a casual observer, CODATA's sponsorship of a specific activity such as the Fundamental Constants Task Group may seem anomalous in the context of a forward-thinking strategically activist body. However, CODATA's support of such activities is important in anchoring the organisation to its scientific bedrock. Furthermore, all such Task Groups are independently reviewed every two years to assess their effectiveness and relevance. Only active, relevant, Task Groups continue to be supported by CODATA. We do not therefore believe that a further review of these is justified.

Recommendation 5. The CODATA Executive should actively commission Working Groups and Task Groups to meet specific objectives detailed in the Strategic Plan, under the guidance of a technical advisory committee and such other advisory bodies as it chooses to constitute. In addition, the existing mechanism for proposing and approving Task Groups should be continued.

3.6 Funding

Projects identified by CODATA as essential for the fulfilment of its mission must be adequately funded. CODATA is very fortunate in being able to draw upon a large amount of voluntary effort by its members in the spirit of scientific collaboration. Nevertheless, funding is necessary for travel, administrative and secretarial support, and to cover the costs of resources involved in publishing the results of its activities. Furthermore, it may be necessary to resource further staff if the scale of strategic activities is to be increased.

Short-term funding to support specific project-driven activities is more easily available than longer-term funding, and we applaud the success of recent CODATA Task Groups in obtaining such funding. Often a project-driven activity in pursuit of CODATA goals will be well-aligned with the goals of another organisation, research fund, or commercial company, and so we recommend that Task Groups and Working Groups actively seek funding for such activities from bodies such as:

- national or regional research and development funds (*e.g.* the UK Joint Information Services Committee, the EU FP6 research framework, NSF, APEC, *etc.*),
- international organisations such as ICSU or UNESCO,
- commercial companies, either through sponsorship or by contractual funding.

Longer-term funding is also important. The day-to-day business of CODATA is currently funded by the contributions of the National Members and other member organisations. This supports the current operations of the Secretariat, activities such as the CODATA Prize and *Data Science Journal*, and seed funding for a modest number of Task Groups. However, it would not permit a significant expansion of CODATA activities, and it is insufficient to support additional long-term projects, such as web portals to act as catalogues of data providers and archives, or a registration agency for digital object identifiers of scientific data sets. These would be worthwhile activities for CODATA, but are not feasible with current resources.

If CODATA wishes to play a leading role in driving the strategic direction of science data management, some modest expansion is likely to be necessary, and this will require additional funding. If some of this expansion included the provision of services (*e.g.* a global science data registry) then it is possible to derive some income from such services. It should be noted that there exist successful models of commercially funded open-access services (*e.g.* Google).

Thus, options for long-term funding include:

- additional national members (*e.g.* few European countries are currently CODATA members)
- subscriptions from companies or organisations that make use of CODATA services.

We recognise the chicken-and-egg nature of these: funding will not be forthcoming unless CODATA provides demonstrable benefit, but without funding such activities are unlikely to start. On the other hand, if CODATA becomes more active in promoting international data strategies, then it is likely to result in additional National Members. We note and applaud the recommendation in PAA #50 that “ICSU should encourage those of its members who are not currently affiliated to CODATA to reconsider this position”, which would be an effective mechanism for catalysing this growth.

This group does not have the resources to explore a business case for the provision of services by CODATA. We are therefore not advocating that such services should be set up, but merely that a qualified group should explore whether this is a feasible way forward.

Recommendation 6. CODATA should actively investigate new sources of funding for both short-term projects and longer-term ventures, and should appoint a small panel of appropriately qualified members to work with the Secretariat to explore alternative funding

sources. This should include examining the feasibility of setting up service provision by CODATA, with an appropriate funding model. It should also ask for ICSU's assistance in recruiting new National Members, as recommended by PAA #50.

3.7 Legal and ethical rights and obligations

Several recent technology-driven developments are forcing a reassessment of intellectual property rights (IPR) legislation and changes in the attitude of society towards IPR. This is having a profound impact upon the handling of scientific data and information. On the one hand, the movement towards open-access publishing is challenging conventional transfer of copyright to publishers. On the other hand, protectionist attitudes towards material stored in commercial databases threaten access to scientific data. The ICSU PAA notes the anomaly that increased commercialisation and profit generation from scientific data is occurring in parallel with pressures to de-commercialise scientific publishing; this may be related to the shifting emphasis on IPR.

Practising scientists often have an incomplete understanding of such issues, and we believe that it would be helpful for an interdisciplinary body such as CODATA to report to the community of science the implications of evolving legal practice in IPR, and to interpret their relevance to the broader goals of science as a communal activity for the public good. We believe that it is also important for CODATA to continue to identify and challenge developments in IPR legislation that threaten scientific research, and we applaud the activities of the CODATA Secretariat and individual National Members in recent years to oppose or constrain poorly-conceived legislation. We note that considerable expertise in this area is already available through the ICSU/CODATA *Ad Hoc* Group on Data and Information.

Part of the problem has been that the scientific community has not, in the past, articulated any broad data management principles or policies, leading to the erroneous view, by non-scientists concerned with IPR, that science has no opinion on how data should be managed, leaving other IPR stakeholders free to impose their views on scientific data. To correct this, a data framework as proposed by PAA is essential.

We note also that other legal rights and ethical issues arise in the context of providing equitable access to scientific data. Examples are privacy rights associated with medical or other human-population data; responsibilities in presenting or interpreting data relating to matters of public concern (such as global warming); and data that might prove useful to terrorist or subversive groups in the furtherance of illegal or violent activities. While many of these will be specific to individual disciplines, they raise complex questions, and there is merit in having an overall framework for the responsible management of rights and ethical imperatives within the scientific endeavour.

Recommendation 7. CODATA should work with the ICSU/CODATA *Ad Hoc* Group on Data and Information to monitor developments in intellectual property rights and other relevant legislation, and provide recommendations for best practice within science. This important issue needs to be addressed within the ICSU framework, and so CODATA should play an active role in developing and promoting this.

3.8 The Digital Divide

The "Digital Divide" refers to the widening gulf between those with high-bandwidth access to information, data, and web services, and those who do not. Those who do not are further disadvantaged by this lack of access, making it even less likely that they will gain access in the future. Often the term is used to refer to the gulf between developing and developed nations, but it can also refer to the poor information services available to indigenous (and often geographically remote) inhabitants of an otherwise affluent developed country.

CODATA currently directs significant attention to Digital Divide issues. Two recent workshops in Brazil and China focused on access to scientific information resources in developing countries, and another workshop is planned for the southern African region in late 2005. In addition, there are three CODATA Task Groups specifically focused on data issues in developing countries – in Asia-Oceanic Countries, in Africa, and in data archiving in developing countries. In the past, CODATA has organised tutorials and educational workshops in developing countries, but we note that few have taken place in the last few years.

Recommendation 8. CODATA should continue and intensify its focus on overcoming the challenges of the Digital Divide, and should set that as a key goal within the CODATA Strategic Plan. We recommend that it should establish a series of regular tutorials and workshops on data management and archiving in developing countries.

A key means of overcoming the Digital Divide is to promote Equitable Access to data and information, and we note that this is also in the interests of the global scientific venture. Recognising the recent OECD and Berlin declarations, and the spirit of the WSIS meetings, CODATA should promote Equitable Access policies as a key part of its strategic plan, with a goal being that publicly funded data should in general be available on an Open Access basis.

We also note the importance of the accessibility of journals in developing countries, and recognise that CODATA shares these goals with INASP and ICSTI. CODATA should work with these organisations with the goal of making all forms of data and information accessible to scientists in developing countries.

Recommendation 9. CODATA should actively promote the principle of Equitable Access to data and information, and in particular promote a principle of Open Access to scientific data obtained with public funding. CODATA should also work with INASP and ICSTI to promote Equitable Access to journals and other published information in developing countries.

3.9 Data Quality

Traditional methods of assessing the quality of scientific data are also coming under challenge from technical developments. Data collection and first-stage processing now often proceed at a rate that outstrips publication through recognised peer-review channels. Some types of open-access publishing mechanisms bypass or defer formal peer review. Open-access technologies make it possible for individual laboratories or research groups to make their data publicly available outside the normal channels.

The ease of transferring and transforming large volumes of data increases the potential for data corruption, whether through inadvertent technical error, misplacement within data collections, or active malpractice. We need to encourage standards for verifying the integrity of data sets at the machine level, to guard against unaudited fragmentation or merging of data sets, and to authenticate the contents of data sets and their provenance. We need also to ensure that data can be deposited in and retrieved from long-term archives with the same level of integrity, complete with the metadata that are essential for correct interpretation of the data.

We have also noted in Section 3.7 that data in the public domain could be misinterpreted or misrepresented to the public (whether deliberately or unwittingly). It is therefore helpful to work towards independent verification of scientific models derived from available data. At a policy level, this may involve raising public awareness of the mechanism of scientific publication and the processes of peer review. In some cases, however, it may also be possible to provide independent technical mechanisms for testing the validity of a scientific model. For example, the International Union of Crystallography operates a public software server that will assess the internal consistency and chemical plausibility of a crystal structure. Of course there are many steps from this to, say, a program that could report (on the basis of community-accepted criteria) on plausible scenarios for global change, given an input set of meteorological data. Nevertheless, CODATA is surely the right body to consider and develop such ideas.

Recommendation 10. Recognising the importance of maintaining data quality, CODATA should participate in the development of standards by the scientific community for the validation and verification of scientific data and publications.

3.10 eScience

eScience is the application of advanced Information and Communications Technology (ICT) to enable major advances in scientific research. It includes high-performance computing techniques to model scientific hypotheses and visualise complex data sets. One goal of eScience is to achieve the same transparency for scientific data as the WWW currently gives us for documents, so that all the world's data are available from your desktop, regardless of location or which side of the Digital Divide you sit. eScience is rapidly growing, and is being embraced by governments and funding agencies as a priority area, because of its potential to increase the cost-effectiveness of many different areas of research.

Fundamental to e-Science is Grid Computing, which is the creation of infrastructures to share data, applications and resources. For example, Information Grids emphasise data access and sharing, typically involving very large data collections and powerful computing resources linked across the globe. Other aspects of eScience produce tools and services to enable scientists to access data and techniques in much more powerful ways than previously possible, thus increasing their scientific productivity.

Many of the issues discussed elsewhere in this report, and in the PAA report, are central to eScience. While CODATA is not primarily an eScience forum, the issues being addressed by CODATA will increasingly be influenced by eScience issues and pressures. CODATA should also recognise that an opportunity exists to build eScience on a solid foundation of good data management principles, and that its contribution to eScience will be welcomed by eScience practitioners.

Recommendation 11. CODATA should actively pursue linkages with global eScience groups, and should offer to co-host eScience and data conferences.

3.11 Specific PAA Recommendations

In Appendix 3 we list specific recommendations from the PAA report that are relevant to CODATA. While many of the key issues have already been addressed by the recommendations above, we list in the Appendix several specific responses that we believe to be appropriate to the specific recommendations.

Recommendation 12. CODATA should respond to the specific PAA recommendations with the responses listed in Appendix 3.

4. Conclusion

The ICSU PAA report presents a vision that is well-aligned with the implicit goals of CODATA. CODATA should embrace the opportunity to make those goals explicit, and work with other ICSU bodies to bring sound data management principles to science. To do so will not only further science, but could re-affirm CODATA as the world’s peak body for the management of scientific data.

Acknowledgements

We thank Kathleen Cass for her help and guidance in assembling this report. Some of the comments in this report have been taken from the US CODATA National Committee response by Julie Esanu and Paul Uhlir, whom we gratefully acknowledge for their contribution.

Appendix 1: A Suggested Outline and Process for the CODATA Strategic Plan

A1.1 Outline

We expect that many of the elements of the Strategic Plan will be identified from CODATA's existing programme of activities, while others will be generated by the CODATA response to the PAA report. Formulation and adoption of the Strategic Plan should be an open process based on the broad consensus of the membership. It is not within our terms of reference to produce a strategic framework, but we outline some examples of the ideas and activities that we would expect within such a strategy document. Many are expressions of activities that CODATA is already undertaking.

1. *Full and equitable access to scientific data*
   Relevant activities may include the propagation of agreed principles (*e.g.* the ICSU/CODATA Principles for dissemination of scientific data); investigation of the role of open-access databases (*e.g.* funding models; their relationships – complementary or competitive – to existing scientific databases in the public and private sectors); provision of reduced-cost or otherwise specially managed access for scientists in developing countries.

2. *Monitoring and development of intellectual property rights to scientific data*
   Recent developments in intellectual property legislation, copyright law, patentability, and duty to disseminate results of publicly-funded research have implications that are not always understood by, or apparent to, the scientific community. Our freedom to create and access public-domain databases has been threatened by poorly-conceived legislation, which ignored scientific needs because these had not been articulated to the wider community. CODATA should explore the proper exercise of rights in intellectual property and privacy within the scientific enterprise, actively participate in the development and promotion of the ICSU data framework, and represent the interests of science in the wider public legislative arena.

3. *Long-term preservation of, and access to, scientific data*
   Action is needed both at the technical level (to define practices for ingest, migration and management of digital information over extended time scales), and at the policy level (to raise awareness of the need for long-term preservation and curation, to require archiving to be stipulated as a condition for research funding). A particular challenge for several disciplines is the digitisation of data that is formally in the public domain but is not yet digitised and is therefore relatively inaccessible.

4. *Interoperability and data storage/exchange standards*
   While many disciplines are actively developing standards for the representation and exchange of data within their own domains of interest, more needs to be done in order to exchange information effectively between different disciplines. Furthermore, it is likely that different disciplines are facing similar challenges and can learn from each other. CODATA should be proactive in catalysing cross-fertilisation between disciplines.

5. *Data quality, validation and authentication*
   New technologies have brought new opportunities to improve the quality of raw and processed data, and the scientific interpretation of models tested by data. Unfortunately, they have also brought fresh opportunities for data corruption, misrepresentation and misinterpretation. CODATA should encourage and promote technical routes towards

improving data quality, as well as establishing ethical and legal guidelines on quality assurance.

6. *Capacity building*
   The “Digital Divide” is a manifestation both of restricted access to data through limited financial or technical resources (covered partly at point 1 above), and of limited human resources. Even within the developed world, care must be taken to ensure that the necessary levels of education and training are available. CODATA has an important role to play in establishing standards of training and education, and in facilitating workshops and education programmes in the developing world.

This list of suggestions is not to be considered exhaustive, but is intended rather to suggest the type of headings under which strategic goals may be classified. Many individual activities will straddle two or more categories, and many of the specific interests of CODATA will overlap with those of other bodies.

A1.2 Mechanism

We have suggested an outline for a strategic framework, but significant work is needed to extend and refine it to the point of a document properly expressing the mission of CODATA. In addition, significant work will be needed to identify and propose specific objectives within each category of activity that are realistically achievable. It is expected that the Executive should appoint a Working Group charged with drawing up the Strategic Plan in consultation with the members of CODATA, representatives of ICSU and its related organisations.

A1.3 Implementation

Overall responsibility for implementation of the Strategic Plan should lie with the Executive, who should report on activities within the framework of the Plan annually to the membership, as well as formally to the General Assembly. It is expected that the Plan will list a number of objectives that CODATA wishes to meet over a particular time span (*e.g.* 5 years). In working towards these objectives, the Executive should appoint or commission Working Groups and Task Groups as appropriate, although independent proposals by the membership for Task Groups will continue to be welcomed. It will, however, be useful to relate the activities of all the CODATA Task Groups to the strategic framework, in order to provide context and a coherent overview of CODATA’s project-based activities.

The Executive may also find it helpful to appoint advisory committees as required (in accordance with By-Law 6.11 of the CODATA Constitution).

Appendix 2: Technical Developments

We suggest here a number of examples that are potentially achievable by, or under the direction of, the resources at CODATA's disposal. As with our suggestions for the strategic framework, these are not intended to be either exhaustive or definitive, but rather to act as a starting point for further thought.

1. *Core metadata for science*
   Within the publishing field, the Dublin Core metadata set of terms is used to identify generic attributes of publications that allow their harvesting, classification and querying by a variety of service providers. This metadata set (in aiming to have universal relevance) is rather weak in describing the attributes of scientific information that would be useful for providing value-added services to the scientific community (*e.g.* the relevant discipline or scientific speciality), and is entirely deficient in describing scientific data sets. Examples of attributes useful in classifying data sets are whether the data refer to time series or individual measurements, what material phenomena they describe (*e.g.* fundamental physical constants, crystal structures, astronomical objects, meteorological data *etc.*). Of course, the characterisation of data in detail must be left to the individual disciplines; however, the definition of a high-level metadata framework (similar to the Dublin Core) for scientific information may be achievable by a panel of cross-disciplinary experts working together with experts in the Dublin Core and other metadata standards.

2. *A registry of identifiers for scientific data sets*
   Another exciting development in the publishing world is the registration of digital object identifiers that are uniquely associated with each published scientific paper or chapter of a book. A registration agency, CrossRef, exists to collect and store these identifiers with metadata describing the "digital objects" they refer to. The result is a resource that provides the ability for scientific publishers to locate and provide links to published papers. A parallel development in the data world (a universal database of scientific data sets) would have the potential to facilitate distributed database querying and interoperability. It is, we believe, unlikely that CODATA itself could act as such a registration agency; but it could play an important role in identifying the need for such a resource and in bringing together scientific database and data storage managers to explore the mechanisms and economic basis for a data DOI registration service.

3. *A catalogue of data resources*
   At a more coarsely-grained level, it would be beneficial for science to have a reasonably complete catalogue of sites holding and distributing scientific data sets. Two possible approaches to this are:
   (i) Delegate to each Scientific Union the role of cataloguing relevant databases in its area of science, and collect together in one location (*e.g.* a CODATA-maintained web site) the results classified by discipline. Although labour-intensive and subject to the varying resources of the individual Unions, the centralised posting of such information may be effective in encouraging contributions.
   (ii) More speculatively, and based on technical developments, data resources could advertise themselves on the web through a standard protocol (a likely candidate at this stage, arising from the Open Archives initiative, is the Protocol for Metadata Harvesting OAI-PMH). Such resources could then be harvested automatically by web robots and catalogued, again by a central agency such as CODATA. Successful operation in this way would depend on

a richer set of relevant metadata, as described in point (1) above. Although the collection of information about data providers would be much facilitated by this technical mechanism, it is likely that some moderator or editor would be necessary to manage the resulting catalogue to the necessary level of quality.

4. *A catalogue of archives*
   In similar spirit to the above, we suggest a catalogue of data providers or managers that satisfy the requirements of an "archive" of scientific data as expressed by the Open Archive Information System (OAIS) reference model. The entries in this catalogue could be a subset of those in point (3) above, but with a public statement of their policies and mechanisms related specifically to long-term preservation. Again, this catalogue could be put together either under the direct responsibility of individual Scientific Unions or assisted by robotic harvesting. Hosting it on the CODATA-sponsored web site would provide each contributing archive with a very public statement of the importance it attached to the preservation process, and could facilitate confederations of archives, leading to mutual backup provision and further cooperation. (We note in this context the pilot CODATA/ICSTI Portal on Permanent Access to Scientific Data and Information that will be presented to the 2004 General Assembly.)

5. *Automated rights management*
   Current developments within the "Creative Commons" movement are providing machine-readable statements of intellectual property rights and licenses for reuse of digital material. This combines the technical requirements of machine-readable rights management (thus facilitating automatic exchange and use of information) with the classification of appropriate rights of fair use. Although this initiative arises largely in the context of open-access developments, it seems to us potentially to provide a mechanism for machine-readable rights management that might be incorporated into the use and re-use of content supplied by commercial and public-sector data providers.

These are provided as examples only, but they illustrate the sort of interdisciplinary activities, many based on technical developments, that should be identified and put forward for consideration within the strategic objectives of CODATA. So far as we aware, they represent new ideas not currently addressed by existing Task Groups (the existence of the Portal on Permanent Access became known to us only as we progressed through the drafting of this report).

Appendix 3: Response to Specific PAA Recommendations

The group suggests that CODATA respond to ICSU concerning the specific recommendations in the PAA report in the ways outlined below. We discuss first the recommendations directed specifically at CODATA; we then identify and respond to a number of general recommendations to ICSU where we feel CODATA has specific expertise or knowledge to offer; and we lastly consider the place for CODATA within ICSU's grand vision. In each case, we have framed the response in such a way that it could be incorporated in a formal memorandum of response to ICSU. We emphasise that these are only suggested responses at this stage, and are contingent on agreement with the other recommendations in this report.

A3.1 Recommendations directed at CODATA

48. CODATA should develop a clear long-term strategy that focuses on key international data management and policy issues and should place a strong emphasis on eliminating the digital divide.

- CODATA plans to:
  - develop and publish a mission statement and Strategic Plan for action on a five-year timescale, and report progress on activities within that strategic framework;
  - strengthen its grasp of technical issues and develop active projects in international and interdisciplinary data management; and
  - establish the elimination of the Digital Divide as a key goal within the CODATA strategic plan.

- CODATA already concentrates on issues relating to the Digital Divide. Two recent workshops were convened in Brazil and China that focused on access to scientific information resources in developing countries, and another related workshop is planned for the southern African region in late 2005. In addition, there are three CODATA Task Groups specifically focused on data issues in developing countries – in Asia-Oceanic Countries, in Africa, and in data archiving in developing countries.

49. The lines of communication between CODATA and ICSU need to be improved. CODATA should continue to develop a closer working relationship with ICSU bodies such as INASP and ICSTI in areas where there are complementarities and clear added value.

- Over the last few years, CODATA has worked to strengthen its relationships with ICSU and its member organisations. The work related to WSIS is one example of opening the lines of communication and collaborating on an issue very important to the international scientific community. CODATA has also worked with ICSTI over the last few years in areas where there is a clear link between digital data and literature issues or activities, such as preserving and accessing scientific data and information. CODATA has collaborated with ICSTI to convene several workshops (Open Access workshop in Paris, 2003; a data and information workshop planned for southern Africa in 2005; and an "InfoCommons" workshop planned for Paris in 2005) on these issues, and will continue to work closely with ICSTI to promote access to scientific information resources. An example of more operational collaboration between CODATA and ICSTI is their current initiative to create an online portal on Permanent Access to Scientific Data and Information Resources, a prototype of which will be demonstrated in Berlin at the upcoming CODATA Conference.

- In addition, CODATA and INASP are working together to help ameliorate the digital divide. INASP recently participated in the CODATA workshop in Beijing that was focused on strategies for preserving and accessing digital scientific information resources in China, and will collaborate with the CODATA archiving task group, the South African and US national committees, and ICSTI on a related workshop looking at sharing and preserving digital scientific information for sustainable development in southern Africa. These activities do not detract from the core missions of these organisations, but rather enhance and strengthen them.
- Identification within the Strategic Plan of specific goals relating to issues of international development, capacity building, or the synergistic relationships between data management and scientific publication will throw into clearer relief the areas where collaboration with other ICSU bodies is most needed.
- A beneficial mode of communication would be an exchange of strategic plans between ICSU, CODATA, and related organisations.
- CODATA's active participation in the proposed Scientific Data and Information Forum (PAA #58) is an opportunity to strengthen its relationship with ICSU.

50. CODATA needs a more inclusive worldwide membership. ICSU should encourage those of its members who are not currently affiliated to CODATA to reconsider this position.

- The National Members of CODATA contribute vitally to its funding and direction, but there may be structural reasons why National Committees are difficult to establish even in countries that are involved heavily in scientific data activities (*e.g.* the lack of a scientific or other coordinating body able to act as a focus for national efforts, or a bias towards one particular model of data management – *i.e.* private or public sector). Nevertheless, CODATA will actively strive, with ICSU's help, to recruit more National Members.
- CODATA also plans to consider alternative ways of engaging national representatives, *e.g.* by increasing the number of Affiliate National Members, by organising and recruiting more member organisations by country, or by developing regional memberships. The support of ICSU, as recommended in this area, would be very helpful, and CODATA will actively seek this assistance from ICSU.

51. While developing its long-term strategy, a short-term CODATA focus on implementation of relevant aspects of the Science in the Information Society *Agenda for Action* and preparation for the World Summit on the Information Society II is both appropriate and valuable.

- CODATA is heavily committed to further work towards the WSIS meeting in Tunis in 2005. Activities include a full-day session at the CODATA conference 2004 in Berlin.
- CODATA's current "focus on the implementation of relevant aspects of the Science in the Information Society Agenda for Action and preparation for the World Summit on the Information Society II" is not short-term. CODATA and its US National Committee has been working on related issues such as data access for a very long time (see, for example, the ICSU/CODATA Principles of the dissemination of scientific data, or the *Bits of Power* report organised by the USNC/CODATA, NRC, 1996). The WSIS work is a continuation of CODATA's work in this area for the past decade and is not a short-term goal.

A3.2 Recommendations to which CODATA can contribute

5. ICSU should work with its members and key partners to establish guidelines on good practice in public sector data management.

- CODATA will actively strive to promote principles of Equitable Access, and will build this in as a key goal of its Strategic Plan, with particular emphasis on Open Access to publicly-funded scientific data. It will highlight and promote existing guidelines (*e.g.* the ICSU/CODATA Principles for dissemination of scientific data, the CODATA/Inter-Union Bioinformatics Group report), and serve as a focus for common principles for data access to be shared across scientific disciplines. CODATA will actively participate in the development and promotion of the ICSU data framework.

14. ICSU should work with its members and relevant bodies in encouraging the coordinated development of digital libraries and their integration with journal publishing and data systems.

- A goal of the CODATA strategic plan will be the removal of obstacles to the transparent passage of data from published journals to databases and data centres. CODATA will work with publishers and INASP to develop guidelines for publishers to facilitate this process.

22. ICSU should work with its members to promote the development and use of flexible, open, and easy to use community standards for metadata. These standards should be interoperable and independent of specific hardware and software platforms. Guidelines for their use should be widely circulated and incorporated into data management training courses.
And:
32. ICSU and its partners should "promote interoperability principles and metadata standards to facilitate cooperation and effective use of collected data and information," as recommended in the Science in the Information Society Agenda for Action.

- We have identified the definition of a high-level set of common metadata as a project suitable for investigation by a CODATA Task Group.

27. ISCU should foster discussion within the scientific community, including its members and interdisciplinary bodies, on criteria, institutional structures, and models for decision making related to the permanent preservation of scientific data and information.

- There is a specific issue around data on non-digital media being transferred to digital form so that it can be made available electronically. Techniques and issues are common to many disciplines, and CODATA is prepared to take a lead in coordinating these activities across disciplines.

33. ICSU should bring together representatives of voluntary data confederations and distributed data systems to discuss what has been learned over the past ten years about what contributes to the success of voluntary data confederations, what undermines them, and what must be done to preserve and enhance access to the data in the future.

- CODATA will offer to collaborate with representatives of other appropriate organisations to organise a conference on this subject.
- CODATA has also identified the use of common high-level metadata and de facto standard protocols for metadata harvesting in facilitating federated and distributed data networks.

35. Emphasis should be placed on the need for professional data and information management in those countries where the scientific data infrastructure is being constructed.

ICSU bodies, such as the International Network for Access to Scientific Publications (INASP) and CODATA, have a key education and training role to play in these countries.

- CODATA sees this as a vital step in overcoming the Digital Divide, and plans to establish tutorials and educational workshops in Developing Countries, and looks forward to working with INASP on this and related issues.

36. ICSU should continue to stand firmly behind the principle of full and open access to scientific data.

- CODATA emphasises its commitment to the unhindered and equitable access to scientific data, both explicitly in its mission statement and through its activities. A goal of its strategic plan will be to disseminate these principles throughout its member organisations.

39. Governments and other bodies concerned with international and national policy development, should ensure that IPR legislation recognizes the value of ensuring full and open access to data for scientific research and education purposes.

- CODATA was active in identifying the deficiencies in the WIPO legislation that was proposed 2-4 years ago. CODATA will explore how best to continue this momentum, perhaps by setting up a Working Group, so that access to scientific data may be safeguarded in future.

A3.3 CODATA and the role of ICSU

1. The ICSU Priority Area Assessment (PAA) Panel on Scientific Data and Information strongly recommends that ICSU assume an international leadership role in identifying and addressing critical policy and management issues related to scientific data and information.

- CODATA is ideally placed to work with ICSU in identifying and addressing significant changes in the management and use of scientific data. In developing its mission statement and Strategic Plan, it will pay particular attention to the major dangers threatening fair and equitable access to data. CODATA is enthusiastic to play a key role in the ICSU process.

58. The panel recommends that, in parallel to the development of the long-term strategic framework, ICSU establish an international Scientific Data and Information Forum (SciDIF) involving all the key stakeholders: ICSU members, interdisciplinary bodies, science funding bodies and other data providers and users. Through SciDIF, ICSU should aim to ensure that the full benefits of new data and information technologies and capabilities are extended to scientists throughout the world

- It would be appropriate for CODATA to play a key role in SciDIF, and CODATA is keen to do so.