CODATA Workshop on Archiving Scientific & Technical (S&T) DATA

20-21 May 2002
Pretoria, South Africa

Organizers: South African National Committee for CODATA, CODATA Working Group on Data Archiving, National Research Foundation

Report

1 Introduction

The 1st CODATA (Committee on Data Archiving) Workshop on Archiving Scientific and Technical Data was held at the National Research Foundation (NRF) in Pretoria, South Africa, on 20-21 May 2002. The workshop was organized and sponsored by the South African National Committee (SANC) for CODATA, the CODATA Data Archiving Working Group, and the NRF. The program for the workshop is included as Annex A.

The workshop brought together more than 70 people representing more than 30 South African science agencies, universities, and archives as well as scientific data archiving experts and CODATA members from Africa, Asia, Europe, and the United States of America.

This report reviews the workshop objectives, summarizes the issues discussed, and briefly outlines some future CODATA events related to archiving S&T data.

2 Workshop Objectives

The principal objective of this workshop was to contribute to a greater understanding of the requirements and major issues regarding S&T data archiving at the SADA and NRF. This included the identification of the roles and expertise needed for the preservation and archiving of S&T data, as well as the underlying conceptual, scientific, technical, management, and policy issues in archiving scientific data. A secondary objective was to make progress on the objectives of the CODATA WG (Working Group) on Data Archiving and to advance preliminary plans for the proposed CODATA TG (Task Group) on the Preservation and Archiving of S&T Data in Developing Countries.

Possible outcomes and deliverables included:

- Identification of science domains for possible inclusion in SADA (South African Data Archive) and NRF data archiving activities;
- Identification, discussion, and analysis of key scientific, technical, management, and policy requirements and issues for expansion of SADA and NRF data archiving activities;
- Drafting of a summary report from the workshop that can be used by SADA and NRF in advancing its broader S&T data archiving plans, and that also can be used as an example of the type of workshop that could be organized by the proposed CODATA TG; and
- Initiating the development of a comprehensive database/directory and e-mail list of S&T data archiving managers and experts from South Africa and the southern African region.

3 Workshop Summary

3.1 Keynote of Dr Khotso Mokhele, President, NRF

Dr Mokhele started by welcoming the attendees and thanking the organizers of the CODATA workshop for their efforts. He emphasized the importance of not only data archiving but data management as well. Dr Mokhele said the lack of data has been found to inhibit the ability of organizations to fully assist clients. And, he added, southern Africa in particular has a poor record of data storage and dissemination.

When it came to power in 1994, the new government was required to draw up legislation and create policies for a new, democratic South Africa. The government had to do so despite a shortage of data and therefore they used their “gut feel” when drawing up policies. This lack of knowledge made the government vulnerable to the influence of outsiders who did have access to data from countries overseas. Attempts were made to impose this outside influence on South Africa.

Members of the country's science and technology community identified this general weakness in the field of data management. The Department of Arts, Culture, Science and Technology (DACST) initiated an audit in 1995-96 on "archivable" data, to assess the S&T system in the country. Due to a number of obstacles present at the time, this audit had only limited success.

At a workshop in Mozambique (1997-98) on land use and land coverage changes, Mozambique lamented the lack of maps and the necessary spatial data on land coverage and climate. This hindered them in their decision making. A few months later at another conference, it was found that South Africa had access to the required data, gathered as part of the CSIR's (Center for Scientific & Industrial Research) defense programme. This is a classic example of the disparity in data collection in southern Africa and highlights the need for a coordinated data archiving and data sharing effort. It also gives an indication of how crucial data is to the developing world.

Data can be very open to interpretation and the need for data integrity is paramount. This is evident in current data about Aids. On a daily basis the media publishes and broadcasts information about Aids. It is one of the greatest challenges that the subcontinent faces. Yet the data on Aids is so contentious that there is considerable (counterproductive) haggling over the results. This demonstrates clearly the need for data integrity.

We hope that this workshop will enable us as South Africa to understand our attitude towards data archiving and management. The NRF recognizes the importance of knowledge management and has seen fit to create a knowledge management post at executive level. This division is 30-40% inward looking and 60-70% outward looking in order to play a role in the national discourse on how knowledge is managed in the national context.

The country has taken too long to see the value of data collection, archiving, storage and processing. People need to embrace the value of data. The NRF must see itself as a champion for the greater national effort. However, this must not be an isolated effort. There must be synergy between various role players.

There have been hindrances to the archiving of data. This includes people's difficulties in associating archiving with the electronic media. The data chain is also compartmentalized i.e. those who archive data are often not those that produce it. The data archivists then have to make sense of the data that they themselves do not produce. Therefore there is a need to construct relationships between those that produce data, those that archive it and those that mine data.

South Africa's S&T organizations are managed according to national and international norms. These organizations are built for science. SADA was constructed with a maintenance crew. It was not constructed to conduct research. Ideas from this workshop would help the NRF and SADA to play a better role in helping South Africa.

In conclusion, it is important to note that there is a CODATA task group that is responsible for looking into data archiving for developing countries. It is important that ICSU (International Council for Science) should remain on this path. The ICSU in its next conference in Brazil needs to look at the developing world archiving problems in more detail.

3.2 Scientific Issues: Session 1[1]

The scientific issues and aspects of archiving S&T data include the discipline specific needs and practices of scientific communities as well as interdisciplinary values and methods.

3.2.1.1 S & T data archiving definition and tasks

- Although open to discussion and interpretation in specific disciplines, data archiving is primarily a program of practices and procedures that support the collection, long term preservation, and low cost access to, and dissemination of, S&T data.
- The tasks of the data archiving include: digitizing data, gathering digitized data into archive collections, describing the collected data to support long term preservation, decreasing the risks of losing data, and providing easy ways to make the data accessible.

3.2.1.2 The role of data archiving in S&T

- Data archiving and the associated data centers need to be part of the day-to-day practice of science. This is particularly important now that much new data is collected and generated digitally.
- Experiences in specific disciplines show that it is important that data managers be accepted by the researchers. In some cases this means that some scientists move from research into data management work.
- Furthermore, it is crucial that disciplines and communities develop agreed upon criteria for data quality that enable data to be shared.

3.2.1.3 Challenges of data archiving practice

- Cost of professional data management is not appreciated
- Innovations in data archiving and management practice and technology is essential to accommodate funding decreases
- Diversity of technology make data archiving practice and deployment more complex
- As more and more people expect access to S&T data, science and data management become more interactive
- Data archives and centers need to provide services that add value to the data collections.

[1] Paper delivered by Dr Marten Gründlingh summarized by Prof. Liu Chuang and Dr W. Anderson

3.2.2 Technical Issues: Session 2 [2]

Technical issues include the challenges and opportunities related to the standards, systems (hardware and software), and solutions that support the operational practices of archiving and preserving access to scientific and technical data and information.

There are three important facts about S&T data holdings related to the technical issues of archiving and preserving access to the data.

- First, the volume of digital data being collected is growing exponentially, the number of scientists and agencies involved in collecting and using digital data is also increasing, and many more scientific disciplines are utilizing and being driven by digital data and technology.
- Second, even though all science is increasingly using, and dependent on, digital data, the differences and diversity of S&T data holdings between industrialized and developing countries has increased tremendously in the past 20 years and it continues to increase.
- Third, digital scientific data and information continues to be held in different data standards and formats, stored in different file and file storage formats and systems, and collected and stored in different languages.

In exploring the technical issues it is useful to look at three different styles of data archiving procedures that represent a range of projects.

- One is the centralized large scale data archiving center; e.g., NOAA/USA (National Oceanic & Atmospheric Administration), and NMC/China (National Meteorological Center), both of which collect meteorological data. These centers have well developed archiving procedures and staff. The main challenges and risks are those of managing the huge volume of data that is continually collected.

- A second style is that associated with smaller scientific research projects such as Land Use/Cover Change in East Asia supported by the Asia Pacific Network. In these projects the scientists are both data creators and data users. Furthermore, research projects are more interested in publishing results than in long term preservation of the research data. These projects usually have no standard archiving procedures or competence and little or no documentation about the data. There is a substantial risk that once the project is completed that the data will be lost.

- A third style is associated with data development and data mining projects such as the gene database development for rice. Scientists in these projects are both data developers and data users. As a result they are attentive to data issues, but often use self-designed databases and archiving systems, and restrict access to the data and its documentation. There is often a big risk that the data collected and generated in these projects is commercialized and removed from the public domain.

[2] Paper delivered by Prof. Liu Chuang summarized by Dr W. Anderson and Dr M Makiwane

Procedures and solutions for managing S&T archiving, ,particularly in developing countries include developing hierarchical guidelines for organizing data and mirror site technology for providing access.

It is possible to develop a hierarchical set of guidelines that start with a master guideline for all archiving styles, and include guidelines at the database level, data network level, and the data center level. At the database level the key issues include identification of valuable data, possibly the work of digitization and cleaning, documentation of the data, and the specifics of software, hardware, and even web page development. At the data network level the issues include identification of archived data, checking the data against intellectual property and data policies, and the specification of data access tools and data browsing and download capabilities. There are opportunities at this level to create "pearls" of individual databases and link them together into data information "necklaces." Finally, at the data center level these issues include those of data storage and security, environments for long term preservation, maintaining links to end user applications, supporting inter-operations among different data centers, and increasing the accessibility of the data, possibly by using mirror sites.

For developing countries, mirror site technology could be an efficient way to make their own high quality data available and to participate with scientific projects in other countries. Two other considerations are important for developing scientific and technical data archives. One is participating with others to develop appropriate data policies for both archiving and access. The critical issue is to avoid losing public access to public domain data and data generated from such data. The other critical activity is building data archiving capability for all projects. This requires training scientists and project leaders to expect and use shared data archiving procedures, and training people who can mentor and train others in developing countries.

Guidelines for S&T data archiving for developing countries must

- accommodate multiple disciplines, a diversity of technologies,
- provide procedures that cover the range from identifying quality data, to providing appropriate accessibility,
- must develop methods for selecting items for archiving,
- specify data formats and media,
- develop procedures for writing glossaries and handbooks, and
- provide demonstration of the applications of the data.

From a technical perspective there are **three conclusions** drawn from this presentation. First is that the three major types of archiving procedures are quite useful is looking at the range of data-driven scientific projects and the associated issues. Second, the multi-disciplinary and long-term characteristics of S&T data archiving need to be acknowledged and addressed. And finally, it is of benefit to industrial and developing countries alike to foster international cooperation in archiving, and preserving access to, S&T data.

3.2.3 Management Issues: Session 3 [3]

3.2.3.1 Managing the design and development of a data warehouse [4]

In considering how the database needs of a large-scale, long term (longitudinal) social science based research project could be met and determining what are the particular management problems associated with such a project, the paper touched on the following key issues need to be addressed:

- Clarification of the characteristics of a research data warehouse in comparison to the 'standard' enterprise data warehouse
- The selection of an appropriate systems development methodology for research data warehouse development
- The identification of various clients and their needs, and the implementation of systems and utilities to meet these needs
- Data integration, interoperability and implications for data warehousing
- Coordination of data delivery by researchers with data warehouse requirements with reference to supporting documentation
- Metadata and data dictionary aspects
- Confidentiality, disclosure and intellectual property issues.

3.2.3.2 Framework for Digital Archiving: OAIS Reference Model [5]

The Open Archival Information System (OAIS) Reference Model (RM) has been developed and is now a standard under the International Organization for Standardization (ISO). It addresses a full range of archival functions, and it is applicable to all long-term archives and those organizations and individuals dealing with information that may need long-term preservation.

The OAIS establishes common terms and concepts for comparing archival concepts and implementations, but it does not specify a particular implementation; it identifies a minimum set of responsibilities that must be discharged for an archive to call itself an OAIS archive; it provides detailed models for archival function and for the information associated with archives, and it also provides perspectives on migration, emulation and interoperability among OAISs.

3.2.3.3 South African Data Archive [6]

The objectives of SADA were briefly outlined as follows:

- Acquire and catalogue survey data and related information
- Preserve data against technological obsolescence and physical damage

[3] Summarized by Prof. Steve Rossouw and Mrs H van der Berg
[4] Andrew Paterson, e-mail: anmpaterson@hsrc.ac.za
[5] Donald Sawyer, donald.sawyer@gsfc.nasa.gov, Lou Reich, louis.i.reich@gsfc.nasa.gov, Thierry Levoir, Thierry.Levoir@cnes.fr
[6] Avinash Chuntharpursat, e-mail: sada@nrf.ac.za

- Provide depositors of data with necessary information to ensure high standards of data documentation
- Re-disseminate information for use by other researchers, re-analysis of data, longitudinal and comparative studies, research training, teaching and policy-making decision purposes
- Formulate policies for the scope and content of data and data presentation
- Promote the optimal use of data
- Establishment of virtual archive
- SADA hopes to reflect the government's core priority themes in its holdings, e.g. HIV/AIDS studies

3.2.3.4 Human Resource Development & Management: A Critical Success Factor in Scientific and Technical Data Archiving? [7]

It is not enough to pontificate about policies and strategies for scientific and technical data archiving at a higher level, unless the people who are going to operationalise those policies and strategies at a lower level have the necessary competencies, knowledge and skills. Archiving of scientific and technical data is not merely a mechanical process but rather a human act. Therefore the contextual realities within which scientific and technical data archiving should occur in South Africa cannot be ignored.

Human resource development and management is indeed regarded as a critical success factor in archiving of S&T data. However, a straw poll on training of archivists in South Africa indicates that very few institutions offer any studies in data archiving.

The ESDIS (Earth Science Data and Information Systems Project) Best practices and Benchmark Report (2001) [8] identified best practices from fifteen external 'world class" data centers. Taking those findings into account:

- the current SADA staff would need to "leapfrog" from the current practices of being a "depository" to fully-fledged data archival best practices.
- Having scientists on-site actively using the data in the archive that will generate active feedback on data quality, metadata quality, data services, and expertise in the data and its use, all of which will help to provide the best possible service to users and best stewardship of the data is a long way off in developing countries. Convincing scientists to deposit data is a full-time job on its own!

Multi-pronged strategies for HRD (Human Resource Development) and capacity building were proposed including:

- On the job training
- Continuous Education by form of Short Courses; study visits; knowledge transfer
- Secondment programmes
- Learnerships
- E-learning

[7] Lulama Makhubela, e-mail: Lulama@nrf.ac.za

[8] ESDIS Data Center: Best Practices And Benchmark Report, *Submitted by:*G. Hunolt, SGT, Inc. A. Booth, SGT, Inc. (September 28, 2001)

For CODATA to make a difference in developing countries, it should not only shape policies or establish structures and systems in archiving S&T data, but also concentrate on developing the human resources that can understand such systems and structures. It is suggested that the proposed Task Group for Preservation and Archiving of Scientific and Technical Data in Developing Countries take this as a priority area.

3.2.3.5 Managing S&T Data: Preservation in the Broader Context [9]

S&T data do not exist in isolation but are part of the broader research context. Likewise, the management and preservation of S&T data must consider the wider geographic, political, and information contexts. Ultimately, the data set may be part of a larger context of research documentation including gray literature, technical reports, journal articles, or web sites. The data may be applicable not only in the local research context but for others world-wide. Connections between preservation activities in research communities and the S&T data community were suggested.

- Consider preservation at all stages of the research process
- Manage various types and formats of research output
- Involve multiple stakeholder communities

The following should be considered for future access of data

- Integration across output types, formats, and stakeholder communities
- Shared standards and best practices -- repository management and workflow, federated archives, metadata, authenticity, interactive formats
- Common understanding of technical strategies and economics
- Agreements on roles and responsibilities
- An integrating framework, for example the OAIS that provides a shared way of communicating and standard terminology

3.2.4 Policy Issues: Session 4 of the Workshop [10]
Summaries of presentations [11]

Motivations of government and not-for-profit scientists in research largely based on: intellectual curiosity, creation of new knowledge, peer recognition, career advancement and promotion of the public interest. The values and goals are best served by: maximum availability and distribution of the research results, at the lowest possible cost, with the fewest restrictions on use, and the promotion of the reuse and integration of the fruits of existing results in new research.

The public domain in S&T databases, and the related policy of full and open access to such resources in the government and academic sectors, reflects these values and serves these goals. The policy of “full and open” access or exchange has been defined as “data and information derived from publicly funded research are made available with a few restrictions as possible, on a non discriminatory basis, for no more than the cost of reproduction and

[9] Gail Hodge, e-mail: Gailhodge@aol.com

[10] Summarized by Dr Paul Uhlir and Mr C Sebego

[11] Paul Uhlir, e-mail: PUhlir@nas.edu , Robyn Arnold, e-mail: robyn@nrf.ac.za, Arno Webb e-mail: and A Paterson, e-mail

distribution (COFUR-Cost Of Fulfilling a User Request).“Public-domain information” may be defined as information sources and types that are either ineligible by law for, or expressly excluded from, private ownership and protection, and that therefore may be made available and used without restriction by the general public.[12]

There are many reasons why governments should not use intellectual property protection for their data and information

- The government needs no legal incentives to create the data and information;
- The taxpayer has paid for the production of a government database;
- Transparency of governance and democratic values are undermined by restricting citizens from access and use;
- Freedom of expression would be compromised;
- Much public government information has public-good characteristics; and
- There are numerous positive externalities-particularly what economists refer to as network effects-that can be realized on an exponential basis through the open dissemination of public data and information on the Internet. [13]

The successful exploitation of new innovation tools such as GIS (Geographic Information System), data mining, data modeling, and the Internet itself depends in large part on the ability to access and use large and diverse amounts of data, at the lowest possible costs and with the fewest restrictions on reuse. Conversely, restrictions on access and use represent a limiting factor in the public research process.

In South Africa, a key element of information law is the right of all citizens of access to any information held by the state, as guaranteed by the new constitution. This right is providing a legal framework by the promotion of the Access to Information Act of 2000. The stated purpose of this act is to foster a culture of transparency and accountability in the public and private bodies by giving effect to the right of access to information, and also to actively promote a society in which the people of South Africa have effective access to information to enable them to more fully exercise and protect all of their rights. The free and open access to factual databases created and held by the government is one essential implementation of both the constitutional and statutory right.

Other legislation that is important for preserving access to public information includes the National Archives Act of South Africa (as revised in 2001) and the Promotion of Administrative Justice Act (Act number 3 of 2000). The objectives and functions of the National Archives are to: Preserve public and non-public records with enduring value for use by the public and the state [Section 3-(a)], and to manage records, ensuring accessibility, assisting and setting standards, and promoting the preservation and use of a national archival heritage.

National economic power is becoming increasingly dependent on the production and exploitation of information resources and less on physical resources and industry. However, the gap in information capabilities between the developed and developing world appears to be widening. Illiteracy, innumeracy, lack of computer skills and connections to the Internet are

[12] Reichman and Uhlir

[13] Ibid.

all major roadblocks in South Africa and in the developing countries to enjoying the full benefits of the information age.

At the National level, government information resources, including scientific data, need to be made as useful as possible and broadly disseminated. At the international level, there needs to be greater open sharing of public information on the Internet to help reduce poverty, support disaster prediction, mitigation and management, and promote education and learning. Scientific data are valuable only if they are used.

3.2.4.1 Some key policy issues that were identified for further consideration by the NRF and SADA include:

- Involvement of the scientific and educational community in the national and international IP (Intellectual Property) law and policy discussions.
- Focus on creating, disseminating, and preserving data resources to address fundamental social and economic problems as identified in the NRF HRD strategy plan and the Science and Technology White Paper.
- Development of comprehensive data management and archiving plans by all research institutions based on a presumption of open access in the public domain
- Building institutional relationships in support of established goals and priorities.
- Making the case for adequate investment in data management and archiving as an essential infrastructure element and as a strategic resource for the national system of innovation, linking to the established national research priorities and goals of sustainable development.

3.2.5 Recommendations for Small Data Centers [14]

- Continued funding: funding shared between organizations will alleviate funding pressures
- Cost efficiency: Center must be resident within a larger, relevant organization, to share staff. Smallest computer possible. If possible, remain with the same DBMS (Database Management System) to avoid costly portage and training.
- Data must be quality controlled and regularly updated
- User friendly access (modern database, www)
- Up-to-date products (color, digital, GIS compatible, multi-relational, etc)
- Variable process levels, from raw to (e.g.) averaged
- Appropriate security (donor specific)
- Not only archiving, but partnership (e.g. data analysis) upstream in value chain

[14] Dr M Gründlingh

3.3 CODATA International Conference and General Assembly

Members[15] of the South African National Committee for CODATA will attend the 23rd CODATA General Assembly, to be held in Montreal, Canada, 3-4 Oct. 2002. [16] A Proposal for a CODATA Task Group on the Preservation and Archiving of S&T Data in Developing Countries will be presented to the General Assembly.

Annex A. Workshop on Archiving Scientific & Technical Data Programme

Workshop on Archiving Scientific & Technical Data

20-21 May 2002, Pretoria South Africa

Venue: National Research Foundation

Organisers: South African National Committee for CODATA, CODATA Working Group on Data Archiving & National Research Foundation

[15] Prof. Steve Rossouw and Dr Lulama Makhubela

[16] The 18th International CODATA Conference - ***"Frontiers of Scientific and Technical Data" -*** takes place 29 September-3 October 2002 at the Hotel Delta Center-Ville, downtown Montreal. This four-day Conference is hosted by the Canadian and US National Committees for CODATA. The programme provides for a round table discussion on data archiving.

PROGRAMME

DAY ONE: 20 May 2002

Programme Director: Mr Robert Kriger

08:00 Registration

09:00 Opening: Prof. Steve Rossouw (South African National Committee of CODATA)

09:15 **Welcome & Keynote:** Dr Khotso Mokhele (President, National Research Foundation NRF)

09:45 Keynote: Archiving issues – brief overview by Dr William Anderson (CODATA Working Group)

10:00 Tea

10:30 **Session 1: Scientific issues: Discipline-specific & interdisciplinary**
Chair: Dr L Makhubela (NRF: RI)

Keynote speaker: Oceanographic data: experience and learning by Dr Marten Gründlingh (SADCO)

11:00 ***Discussion***
Rapporteur: Mr Avinash Chuntharpursat (NRF: RI)

11:30 **Session 2: Technical issues**
Keynote speaker: Prof. Liu Chuang (Chinese Academy of Sciences)

12:00 ***Discussion***
Rapporteur: Mr Themba Mohoto (University of the Witwatersrand. Reproductive Health Research Unit)

12:30 Lunch

13:30 **Session 3: Management issues: Lack of standards, metadata; interoperability; compatibility; supporting documentation, etc. pertaining to management not technical**
Chair: Avinash Chuntharpursat (NRF:RI)

Keynote speaker: Managing the design and development of a data warehouse: a case study of the HSRC's Human Resources Development data warehouse project by Dr Andrew Paterson (Human Sciences Research Council)

14:00 ***Discussion***

14:30 Keynote speaker: Framework for digital archiving: OAIS Reference Model, by Donald Sawyer, Lou Reich, Thierry Levoir (French Space Agency (CNES)

15:00 *Discussion*
Rapporteur: **Mrs Henda van der Berg (NRF: RI)**

15:30 Tea

16:00 SADA Demonstration by the SADA project team (Dr Monde Makiwane) (SADA, NRF)
16:30 Closure
Break
18:30 Reception

DAY TWO: 21 May 2002

Programme Director: Tselane Morolo

09:00 **Session 3: Management issues (Continued)**
Chair: Bill Anderson

Keynote speaker: Managing S&T Data: Preservation in the Broader Context by Gail Hodge (Information International Associates, Inc.)

09:30 ***Discussion***

10:15 Tea

10:45 **Session 3: Management issues (Continued)**
Keynote speaker: Human resources management: a critical success factor in archiving S&T data by Dr L Makhubela (Manager: Research Information, NRF)

11:20 ***Discussion***
Rapporteur: Dr Monde Makiwane (NRF: RI)

12:00 Lunch

13:00 **Session 4: Policy issues**
Chair: Dr William Blankley (NRF: Strategic Advice)

Keynote speaker: Policy Considerations for the Management of Public Data Archives by Dr Paul Uhlir (US National Academy of Sciences)

13:30 ***Discussion***

14:00 Keynote speaker: The legislative environment for public record keeping in South Africa by Mrs Robyn Arnold (NRF: Strategic Advice)

14:30 ***Discussion***
Rapporteur: Hettie Terblanche (NRF: RI)

14:45 Tea

15:00 **Session 4: Policy issues (Continued)**

Keynote speaker: Data Wars: Government and global S&T data on the Internet by Dr A Paterson (DACST)

15:30 ***Discussion***

16:00 **Closure:** Prof Steve Rossouw and Dr Lulama Makhubela