Global Terrestrial Observing System

Global Terrestrial Observing System
Data and Information Plan

October 1998

GTOS-18

Foreword

The Global Terrestrial Observing System, established in 1996, is one of three global observing systems working to better understand and address global change issues. The central mission of GTOS is stated as:

> *"to provide policy makers, resource managers and researchers with access to the data they need to detect, quantify, locate, understand and warn of changes (especially reductions) in the capacity of terrestrial ecosystems to support sustainable development".*

Clearly management of data and information, from data collection to dissemination of information for decision making, is a key element of the GTOS programme.

The development of a Data and Information Plan was identified as a priority activity in the GTOS Implementation Plan (GTOS-17). This document has been prepared in response to that and builds on the overall framework laid out for GTOS implementation. It defines the context and confirms the principles to guide GTOS data and information management. It identifies elements of what could become policies and presents an implementation path, including proposed actions and processes required to move forward for effective data and information management within GTOS and externally. This is only a beginning; the details of procedures and guidelines must be developed, augmented and modified through practical application, i.e. they will evolve with experience. Specifically, GT-Net activities present an opportunity over the next 1-2 years, to evolve these policies and procedures in light of operational experience.

Suggestions and comments on all aspects of GTOS Data and Information Management are welcome, both from current GTOS partners and other interested parties. These should be addressed to the Secretariat (GTOS Secretariat, c/o FAO/SDRN, Viale delle Terme di Caracalla, 00100 Rome, Italy - Tel: 0039-06-57053450 - Fax: 0039-06-57053369 – E-mail: GTOS@fao.org).

Gwynneth M. Martin
Principal, The ORBIS Institute

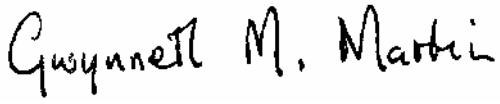

Table of Contents

1. Introduction

1.1 Background

1.1.1 The Global Terrestrial Observing System (GTOS)

There are now three "global observing systems" that, although separate, derive from much the same heritage and share some common goals and principles. Together they provide a means of obtaining a total global picture of the state of the environment of the planet Earth. The first, the Global Climate Observing System (GCOS), deals with climate and climate related aspects of the global environment. The second, the Global Ocean Observing System (GOOS), deals with all ocean and coastal aspects. The third, the Global Terrestrial Observing System (GTOS), deals with terrestrial aspects, including both natural and managed ecosystems. GTOS is the most recently established of the three, sponsored by the Food and Agriculture Organization of the United Nations (FAO), the International Council of Scientific Unions (ICSU), United Nations Educational, Scientific and Cultural Organisation (UNESCO), United Nations Environment Programme (UNEP) and the World Meteorological Organisation (WMO) and supported by a Secretariat located in FAO offices in Rome.

The overall GTOS Implementation Plan (1998) lays out the elements of the GTOS programme and assigns levels of priority to the activities to be undertaken. The management of data and information, from data collection through to the production of information products for policy and decision making, is clearly a key element of GTOS and, in response to this imperative, there are a number of activities identified as "very important to be acted upon as quickly as possible". These include the development of a GTOS Data and Information Plan, and the development and implementation of a number of data-related policies. This document is a response to those priorities.

1.1.2 Joint Data and Information Panel (JDIMP)

In 1994, the Joint Scientific and Technical Committee (JSTC) of GCOS established a Data and Information Management Panel (DIMP) to formulate, implement and oversee the GCOS data and information management systems. Recognising the need for a unified approach to cross-cutting issues, the steering committees of the three observing systems broadened the remit of the Panel to be a joint body - the Joint Data and Information Panel (JDIMP) - with active participation of representatives of GOOS and GTOS. The Joint Panel met first in 1997 and concluded that, although there are clearly useful roles for a single overarching Panel, many data and information management issues at a technical level are specific to the type of observations being made and nature of the phenomena being observed. Thus the transition to a Joint Panel is not automatic and considerable revision of materials developed to date is needed.

Specifically, although the original GCOS DIMP had drafted a GCOS Data and Information Management Plan, that cannot be simply extended to serve GOOS and GTOS. The production of an overall plan for the three systems has now begun, but with every expectation that this will establish general philosophy and common principles, and recognise that there must be individual plans for each Observing System, providing extensions and additions relevant to their differing requirements.

1.2 Purpose and Scope

This document defines the context and confirms the overall principles which serve to guide GTOS data and information management. It identifies proposed policies and, further, presents an implementation path, including proposed actions and processes required to move towards practical operations in the future.

For each major element of the Data and Information Plan the actions required are defined. These are brought together in an integrated plan for implementation that lays out a recommended approach and defines specific tasks required.

It should be noted that the principles and policies identified will evolve relatively slowly over time, requiring infrequent update as conditions change, whereas the work programme elements will change more quickly as implementation proceeds. It is expected that the policy content of this document and the guidelines developed through recommended actions will eventually be incorporated into a GTOS Data and Information Management Handbook, separate from frequently up-dated action plans.

This Data and Information Management Plan is limited in scope to GTOS, although it is prepared in the context of the broader JDIMP planning.

1.3 Strategic Guiding Principles

The GTOS Plan contains a series of Guiding Principles, six of which relate to data and information management. They are:

1 *Data products produced through GTOS have enhanced value to users because they result from related focused inputs from more than one GTOS participating network.*

2 *Data produced through GTOS are high quality, compatible, comparable, reliable and can be used with confidence by policy-making and decision-making bodies, international programmes, and the scientific community.*

3 *GTOS aims to distribute data in an unrestricted manner to encourage free flow of data and information between GTOS data providers, GTOS data processors, and GTOS data users, while respecting the rights of sovereign nations in this regard.*

4 *Recognising that a diversity of data collection methods are available, wherever possible suitable, practical, proven and cost-effective methods for data acquisition, data management, and data analysis are encouraged by GTOS.*

5 *GTOS data are of selected variables chosen to meet specific user needs, and are made available to users in forms, including transformations and derivations, appropriate to those needs.*

6 *On-going GTOS programme actions are taken to help close known data and information gaps relevant to the priority areas of GTOS.*

These principles at the strategic level set the stage upon which more specific principles and policies can be developed as a basis for a Data and Information Management Plan. These are elaborated in the subsequent sections.

2. Approach

2.1 Principles

The terrestrial ecosystems of the World have been observed and studied for centuries, resulting in massive quantities of data and information in a range of formats and media. It is also true that GTOS has, and will always have, limited resources available to perform its functions. These two realities mean that GTOS must carefully circumscribe its scope, and be ruthlessly practical in approaching data and information management. This leads to establishing a series of practical principles to guide policy and actions.

2.1.1 GTOS Facilitates Information Exchange and Use

GTOS is a facilitator, not a collector or "owner" of data. It serves to facilitate and co-ordinate the collection, exchange, processing, integration, and archiving of relevant data, and to promote the generation and use of the data and information products.

2.1.2 Build on Existing Networks and Procedures

GTOS must rely on building co-operative partnerships to achieve its objectives in all areas, including those relating to data and information. Existing institutions and networks will have data and information management policies and procedures in place and it is most unlikely that these will be easily put aside in favour of any directives imposed by GTOS. The approach has to be of building on existing data and information management practices by ensuring that their practices are compatible with GTOS aims. In some instances, existing practices may be deficient and partnership with GTOS will stimulate improvement to the advantage of the new GTOS participant. In other cases, practices will be very much in line with GTOS principles and may have features which could be

adopted to lead to improvement of GTOS overall. Thus while GTOS principles and high-level policies should be established and agreed upon, the implementation will use evolving guidelines rather than a rigid framework.

2.1.3 Build on Existing Institutions

GTOS does not seek to be a new and separate agency, but rather to form linkages with existing data repositories and centres of expertise in information processing and archiving. This means, more specifically that no GTOS "Computer Centre" is to be considered. Not only is it unlikely to be cost-effective, but modern methods of electronic data exchange make it unnecessary.

2.1.4 Use Appropriate Technology

GTOS should employ commonly accepted "good practices" in information management, using, as far as possible, internationally accepted standards for both data and systems technology. This means that GTOS will not seek to be "on the cutting edge" of information technology, but would follow the cutting edge by applying proven widely-used technology, particularly taking advantage of methods to exchange information between disparate systems brought about by the advent of Internet and telecommunications standard interchange protocols and formats.

2.1.5 Allow for Evolution

GTOS data and information management guidelines and related action plans must be subject to frequent review for appropriateness and thereby continuously evolve to stay in harmony with changing overall GTOS plans and policies, and with advancing technology.

2.2 Conceptual Framework

2.2.1 Data and Information

Information scientists often make a clear distinction between "data" (facts that result from measurements or observations of a phenomenon) and "information" (derived from data through assembly, analysis, interpretation or summarisation into a meaningful form). In day-to-day usage the distinction is much less clear. In the context of information systems it is common to use "data" for the **input** to any process and call the output "information" - which may then subsequently be the "data" that is input into the next process and so on. One agency's information (or "information product") is another's data, even though it may be far removed from the initial raw scientific measurement.

Figure 1 illustrates this, with data at the base of the triangle and, moving towards the apex, information is generated from data as they are processed, manipulated,

summarised, etc. At any level, do you have data or information? The figure also illustrates that in moving "up" the triangle -

i) the data (or information) volume is likely to decrease
ii) the nature of the user will change
iii) subjectivity increases
iv) it will take time and resources to move from data to information.

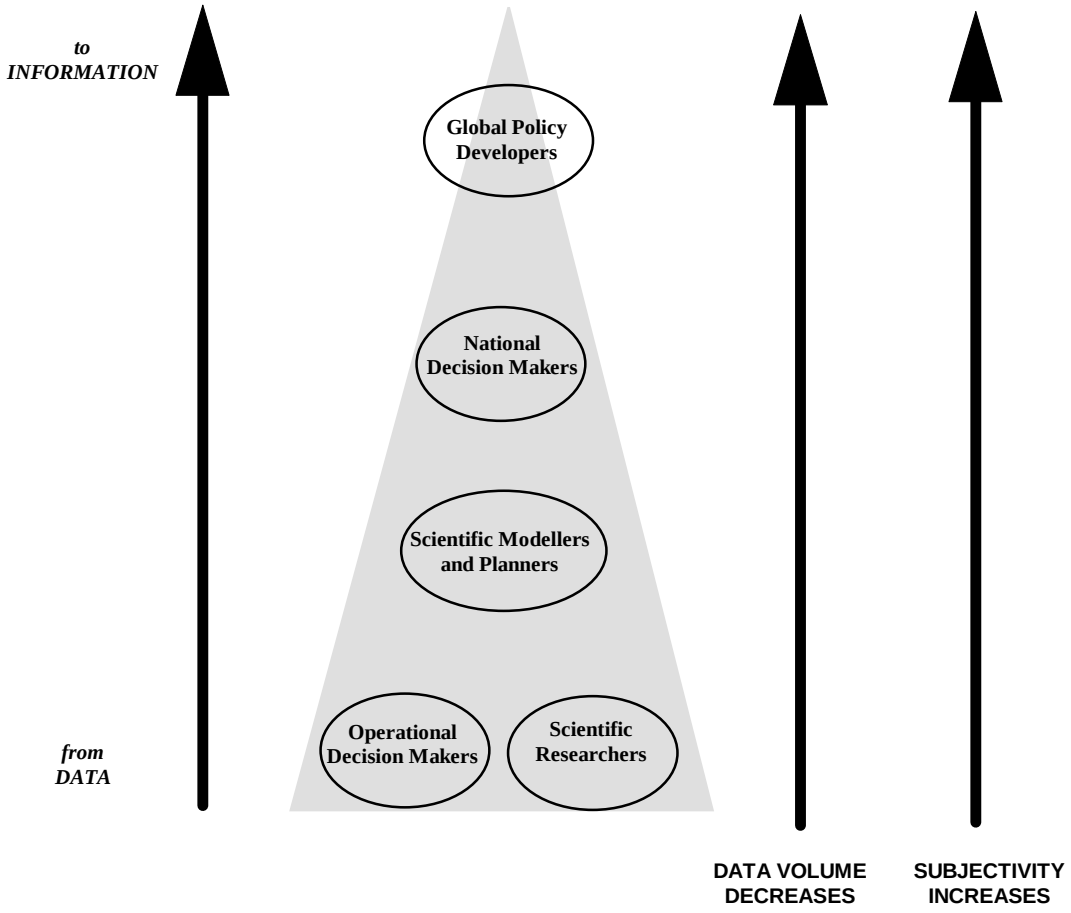

Figure 1

One of the main information management functions of GTOS is to facilitate the move up the triangle - to help transform raw scientific observation into global syntheses suitable for decision making and early-warning.

2.2.2 GTOS Products

The GTOS Plan states that:

> *outputs will range from basic data sets of the variables observed, through transformed, derived and generated data that are in forms more easily handled by*

specific user groups, to technical assessments of the state and trend of particular environmental factors, situations or areas.

In essence, this means that GTOS products may come from any point in the path shown in Figure 1, i.e. a product may be a primary dataset (at the base of the triangle) or some output developed from such datasets.

Given that GTOS is envisaged as a facilitator, all products will come as a result of co-ordination and promotion of activities among GTOS partners. There needs to be clear agreement between all participants as to what will be designated GTOS products and what such designation involves. It is not intended to usurp ownership in any way - that will remain with the appropriate GTOS partner and will be acknowledged as such.

The attaching of a GTOS "label" to a product must imply that the product has been judged to meet GTOS needs and, in the case of datasets, that they are in fact "high-quality, compatible, comparable, reliable". The success of GTOS will depend on products meeting these high standards and therefore some formal procedure to ensure that is achieved needs to be established (see Sec 8, and Appendix 1). This is envisaged to be along the lines of a "sign-off" mechanism through the Secretariat.

2.2.3 Data Management Framework

In moving towards defining its own role, the JDIMP listed the following elements in an overall view of the "end-to-end" data management process for the three Observing Systems, from defining needs through scientific and political priorities to long-term delivery of products. **Science issues** and **political issues** are the driving forces behind the required **applications** and define which **variables** are of concern to the three Observing Systems. **Measurements** are made i.e. new data collected through the definition of observing procedures, the types of instruments, required quality control and metadata. **Data archaeology** is undertaken i.e. existing data reviewed for its utility. Data is **assembled** and **integrated** into a database (data processed, additional metadata provided, quality control and datasets from separate sources merged). **Products** are generated. Data and products are **distributed** to users. **Archive** procedures are undertaken to preserve the various levels of data and information (with the required metadata) for future use. **Metadata** products, such as data inventories, are also generated.

The Joint Panel then identified those elements for which they - a Data and Information Management body - would take primary responsibility, and those for which the Data and Information aspects were more of a secondary, and supportive, nature. For instance, archiving procedures are seen as being a clear Data and Information responsibility, whereas defining overall requirements (issues, applications, variables to be measured) is the responsibility of the various Science Panels. In almost all cases however, it is important to have both scientific and information management expert input to ensure feasible and practical approaches to systems.

Figure 2 focuses on those elements of the end-to-end process most relevant to the content of this GTOS Plan and which will be the main issues to be kept under consideration by a

Data and Information Working Group and/or Panel. The figure shows that archiving, quality management and data and information distribution are activities that may occur at any point in the process. It should also be noted that metadata, mentioned above but not explicit in the figure, is a vital element of the data management framework.

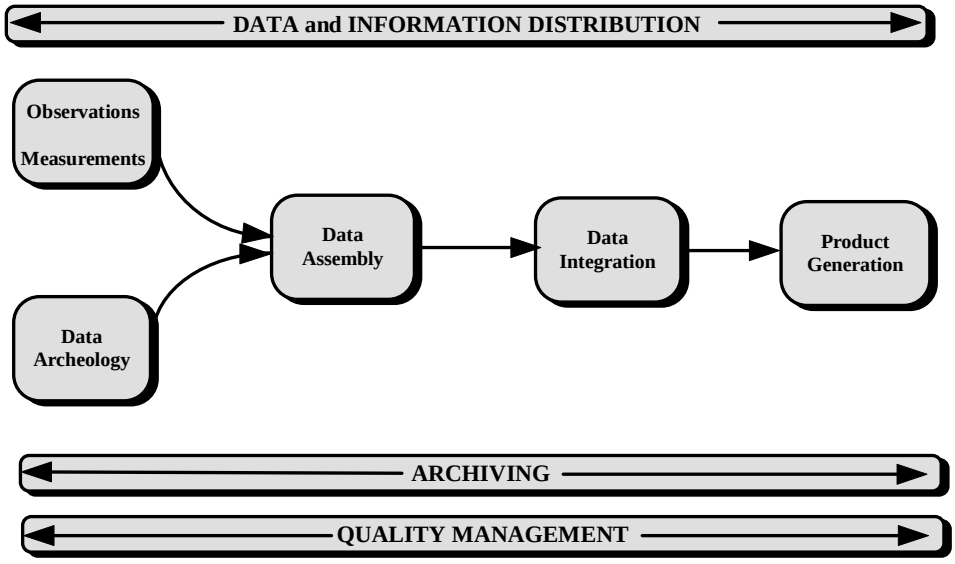

Figure 2

2.3 Structure of the Data and Information Plan

The GTOS Data and Information Plan is structured into three components - process elements, GTOS data management activities, and GTOS support.

The first, **process elements**, addresses the mechanisms required to enable implementation of the Plan, outlining what is in place now, discussing requirements and proposing actions to be taken (Section 3).

The second, **GTOS data management**, covers the range of activities, functions and processes required for the management of GTOS data and information, in the framework described above. It is divided into a number of basic elements, not totally independent, but intended to provide a convenient structure for presentation and discussion. These are:

- User Requirements for Data and Information
- Custodianship of Data and Information
- Access and Release of Data and Information Products
- Metadata
- Data Quality
- Data Harmonisation
- Archiving

For each, there is a brief statement of the recommended GTOS policy, followed by discussion of the issues, and list of actions needed for implementation (Sections 4-10).

The third, **GTOS support**, deals with data management activities which GTOS must undertake as part of its own operations e.g. management of the TEMS database (Section 11).

All the action items identified in Sections 3 to 11 are brought together in an implementation plan that specifies detailed tasks to be undertaken (see Section 12). It is expected that the principles and policies identified will evolve over time, requiring some update as conditions change. This evolution will be, however, fairly slow - perhaps following a review every three years. On the other hand, the tasks required will change more quickly as implementation proceeds. It is expected that the policy and guidelines resulting from completed action items will eventually be incorporated into a GTOS Data and Information Management Handbook. The evolving work programme forms part of the overall GTOS action plan.

3. Process Elements

3.1 Current Situation

The GTOS Plan outlines the bodies and mechanisms through which the various aspects of the GTOS programme will be implemented. Those which have a major role in implementation of data and information management policies and procedures include the following:

1. GTOSSC - The GTOS Steering Committee provides overall guidance to GTOS in all matters. In addition, members of the Steering Committee will play significant roles in activities of the various Panel and Working Groups (see below).

2. GTOS Secretariat - The Secretariat acts as the functional centre of GTOS and will have the responsibility for ensuring that the actions outlined in this Plan (once approved by the SC) are undertaken.

3. Panels:

i) The Joint Data and Information Management Panel addresses issues common to the three observing systems.

ii) The (GTOS) Data and Information Management Panel - This is proposed in the GTOS Plan (1998) as an Advisory Panel to be established.

iii) The Panel on Data and Information Release - This is proposed in the GTOS Plan (1998) as an Advisory Panel to be established.

4. GTOS Data Centres (DCs) - These will be identified as Institutions where GTOS data and information products will be managed, i.e. assembled, verified, catalogued, processed, distributed, etc. They will be existing data and research centres with whom GTOS has established an agreement on their taking on such a role. It is anticipated that, at least initially, they will be centres associated with networks forming GT-Net. (Note that in the first version of the Implementation Plan the term *Data and Analysis Centres* was used. At the Steering Committee meeting in June 1998, that was modified to become simply GTOS *Data Centres*.)

5. Bodies with common interests e.g. FAO, ICSU, UNEP, UNESCO, WMO, IGBP-DIS, WGISS.

Note that all of the above must be co-ordinated and linked with the Science Panels who will be defining and refining the data content and scientific methodologies that drive data and information activities.

3.2 Discussion

Some of the elements listed above are already established with clearly defined roles and scope. The Steering Committee is intended to function as a high-level body, setting overall direction. The Secretariat will have continuing responsibility for the implementation of the approved components of the Plan and for application of developed policies and procedures. The Joint Data and Information Panel, with continued representation from GTOS, should confirm overall principles and plans for G3OS, and provides a forum to identify common areas of interest. Note that this includes the development and functioning of the GOS Information Center (GOSIC), currently being implemented with focus on climate issues, but looking to incorporate GOOS and GTOS interests to develop a true G3OS Center.

However the identification of other GTOS Panels and Working Groups, and the designation of GTOS DCs is at the initial stages and needs firmer definition of scope and mandate.

GTOS Panels and Working Groups

GTOS Panels are expert groups appointed by the Chair of the Steering Committee to consider and make recommendations on designated specific topics. They are composed of both outside experts and Steering Committee members. Working Groups are more informal subsidiary bodies of the Steering Committee. The creation of a Data and Information Management Panel is not only proposed in the GTOS Plan but is identified as a very important activity to be acted upon quickly. Such a Panel would take responsibility for the further development and evolution of the Data and Information Plan (this document). Recognising the constraints under which GTOS is operating, the formation of an interim Working Group would facilitate moving towards implementation (see Sec 12).

The proposed Data and Information Release Panel is described as having (at least) the on-going task of considering how GTOS data and information could be used by countries for

national policy making and economic development. This relates to the continuing identification of user needs (see Sec 4) and, although data and information elements are to be considered, they are not primary (as might be interpreted by the name) and so this is not an element described herein. Renaming this Panel would assist in clarification of roles.

GTOS Data Centres

As noted above, GTOS Data Centres will be existing data and research centres which are formally identified with GTOS with regard to some aspects of the GTOS end-to-end data and information management paradigm. Current considerations of implementation options clearly define several different types of centre functions - for instance, those relating to archival requirements and those relating to distribution. Centres may be involved in the first level of verification of data but not necessarily have facilities for, nor take part in, further data analyses; they may have limited facilities for electronic distribution of data and/or archiving. This will depend largely on the nature and domain of the observational data and the analyses required, as well as the organisational structure and the capabilities of the Centre. Data Centre functions may include one or more of: data collection, data assembly, integration, analysis, product generation, distribution, and archiving. In general, each Centre would take responsibilities within the scope of its existing functions. Some may represent "national assembly centres" as defined in the JDIMP Plan, but in the GTOS context, Data Centres are likely to be more variable in nature and specialisation. It may therefore be desirable to designate different types of DC and specify how they should link into a cohesive whole. Possible categories of Centres are suggested in Appendix 3.

Whatever overall structure is targeted, criteria must be developed for selecting and designating GTOS Data Centres of the various types. These will be used as a basis for negotiation of the role of each potential Centre, and possibly the network partners involved. These criteria will include agreement with the various aspects of GTOS Data and Information Policies as laid out in this and associated documents (see Appendix 3).

To assist both the Secretariat in co-ordinating Centres and GTOS partners in locating data and expertise, information on DCs should be maintained in an online Centres Directory (see Sec 11).

3.3 Actions

- Continue GTOS representation on JDIMP, at a level to allow full participation
- Create a GTOS Data and Information Management Panel
- Assess the needs and benefits for designation of different categories of DC
- Develop criteria for designation of DCs
- Develop linkages with Science Panels

4. User Requirements for Data and Information

4.1 Policy

- GTOS datasets should be collected, developed and managed to meet the known, inferred and predicted needs of the GTOS user communities
- A user needs analysis should precede any major programme of data collection

4.2 Discussion

GTOS aims to supply data and information to a spectrum of users - from scientific researchers to national policy makers. Five global change issues have been defined as priority areas to be addressed. These are:

- changes in land quality;
- availability of freshwater resources;
- loss of biodiversity;
- pollution and toxicity;
- climate change.

The requirement to identify the needs of the different types of user for each of these issues has been emphasised in the GTOS Plan and will be a continuing process as GTOS develops. There will be new users, new issues, and experience also indicates that users' requirements change with time. This is a core activity for GTOS.

The results, i.e. the definition of user needs, must drive data collection and information production, and so must address the basic question: what data/information does the user need? Thus the outputs of users needs analyses should include specifications for the information infrastructure including the following:

- the datasets required
- the analyses to be carried out
- the (information) products to be generated.

Looking at user needs for the five priority global change issues will likely point to the need for a number of "core" datasets - those that are fundamental to any assessment or on-going monitoring. It is these core datasets which should be actively assembled, harmonised and integrated, ready for multiple shared use throughout the GTOS partners.

4.3 Actions

- ensure analysis of user needs clearly identifies data and information requirements
- review outcomes of user needs analyses to identify "core dataset" requirements

5. Custodianship

5.1 Policy

As part of the end-to-end information management framework, all GTOS datasets will have a designated data custodian

5.2 Discussion

Datasets are not generated in isolation. They are developed within organisations and institutions to meet corporate objectives. A custodian is the body **responsible** for the development, maintenance and quality of a dataset, and for the arrangements for access to it. The most important aspect of a custodian is that they should have the scientific and technical knowledge and expertise to be in the best position to assess and ensure data quality, and indicate the appropriate uses and limitations of the data.

The dataset custodian is responsible for:

- building the dataset (primary or secondary data entry)
- maintaining the dataset (up-date and additions)
- ensuring quality (data validation, error correction)
- ensuring integrity (back-up and security precautions)
- providing (and controlling) access
- providing directory level metadata to a central or shared metadatabase
- maintaining full and accurate additional data needed to correctly use the data
- providing advice on appropriate uses and interpretation of the dataset
- providing for the archiving of the dataset

"Controlling access" may include limiting access to only authorised users, and/or imposing usage charges.

Users in turn have the responsibilities to:

- use the data for appropriate applications
- observe any conditions on use agreed with the custodian
- report observed errors and problems to the custodian
- not re-distribute the dataset without the authority to do so
- acknowledge the original source(s)
- not maintain modified copies or alternative "versions" of the dataset

The custodianship concept is designed to ensure the availability of the highest quality data, while reducing redundancy of data collection and maintenance - such as the maintenance of multiple copies or various "versions" of the same data. It thus serves the dual purpose of promoting quality and cost-efficiency.

The most valuable aspect of designated custodianship is clarity of roles - especially with regard to data input and update.

Some form of active "management" of the custodianship functions must be put in place, including the identification and designation of custodians, and monitoring of their conforming to the agreed principles.

It is not unusual to find access controls where the custodial organisation (or groups or individuals within it) have reservations about making the data widely available. However the basic principles and purpose of GTOS make it imperative that custodians do not impose unnecessary conditions to inhibit open exchange i.e. they abide by the GTOS Policy on Data Access and Release (see 6.1).

A custodian may choose to delegate some of the specified functions - such as archiving or day-to-day operational management of the data - but the custodian must maintain the responsibility for these activities by setting standards and monitoring compliance.

Custodian roles within GTOS will be undertaken by at least some of the proposed Data Centres (though it should be noted that carrying out analysis is not a basic function of a custodian). As discussed above (see 3.2), such Centres will have varying levels of capability and facilities and so delegation may be necessary. For instance, a Centre may not operate a shared metadatabase or provide full archiving functions but would delegate responsibility to one which does have the capability in question. This is in line with the possibility of designating different categories of DC.

5.3 Actions

- Develop detailed minimum requirements for a GTOS dataset custodian in the form of a proforma custodianship agreement
- Ensure the defined core datasets have identified custodians
- Establish procedures for assigning, managing and reviewing custodians.

6. Data Access and Release

6.1 Policy

- GTOS data and information should be made available in timely and unrestricted fashion at zero (or minimum) cost
- GTOS data and information should be easily accessible in a variety of forms to meet the requirements of the user community.

6.2 Discussion

Existing organisations and institutions who enter into partnership with GTOS can be regarded as a loosely coupled distribution system through which GTOS data and information can be made available to the user community. At the first level, users must be able to find out what data are available. This can be done using metadatabases and/or data catalogues accessed either electronically, for example through the GOS Information Centre, or using hard-copy publications. The next level involves finding out more detail of potentially useful datasets through examination of the associated metadata. Finally, if a dataset appears to be suitable for intended use, then a simple order-and-delivery process should be available for use. All the steps may be done electronically, either on-line or off-line, or through catalogues and conventional communications. The methods used are likely to be dictated by the user's facilities, i.e. GTOS should ensure that the services can be provided to meet a spectrum of user needs.

Ideally GTOS aims to provide data and information to potential users in an unrestricted fashion. It is not difficult to imagine situations when a data provider may restrict access - for instance, a national government may wish to limit access to data on renewable natural resources. GTOS should then enter into discussion to develop appropriate control mechanisms for data release.

Similarly, GTOS aims to provide data and information free of charge to the user community. It is common practice for data providers to charge the marginal cost of supply - often the cost of generating the copy for the user. This is not unreasonable and such amounts will not usually restrict access. However, data owners may also see their data as a major asset from which they can generate substantial revenue i.e. the price may be high and access would be restricted. In such cases, GTOS will need to enter into negotiations with data owners and the possibility of bi-lateral agreements in some situations (when the dataset in question is clearly a priority) may have to be considered.

Users should, as a matter of course, routinely acknowledge the source of data and information. GTOS should strongly encourage this practice.

6.3 Actions

- Finalise and publicise GTOS data access and release policy
- Build pro-forma agreements to use as a starting point for any necessary bi-lateral negotiations
- Develop guidelines for metadata requirements to enable user browsing and ordering
- Define user needs for mechanisms and tools for browsing and ordering
- Ensure DCs meet user needs for browsing and ordering

7. Metadata

7.1 Policy

- All GTOS data and information must have directory level metadata in accordance with GTOS guidelines
- All GTOS datasets must have adequate metadata at the dataset level.

7.2 Discussion

Metadata[1] are "data about data", describing such things as the location, sources, general content, condition, format, etc. of existing datasets. They constitute documentation covering all aspects of the end-to-end data management process. Metadatabase systems are systems specifically designed to manage metadata i.e. to provide facilities for input, update, retrieval and reporting of data about data. Such systems may be used within a single institution to organise and maintain their own data holdings. They are also used on a broader level and can then provide a mechanism through which data producers can ensure that potential users are made aware of existing data, their nature, and how they might be obtained.

In general, metadata are at two levels. The first, referred to as “directory level”, identifies the dataset through such items as a general description (subject, geographic coverage, dates, collection methods, processing done,...), details of availability (access conditions, costs,...), contact point (for further information and/or ordering). These are items which are essentially common to all types of dataset, regardless of the subject matter. The second or “dataset level”, is subject matter specific, for instance, instrument settings, calibration data, adjustment factors, classification systems and legends, reference standards, taxonomies, etc.

In summary, directory level metadata enables a potential user to judge whether a dataset might be useful for the intended purpose and how to obtain it; the dataset level metadata allows the data to be used correctly, once obtained.

Metadata must exist to enable GTOS to deliver the desired high-quality data and information products to users. They are used specifically to facilitate data access and release (see 6.2), to enable quality assessments to be made (see 8.2), and in archiving (10.2).

Considerable work has been done relating to metadata at the directory level and, as in other areas, GTOS should build on what is currently in use. There is no single accepted global standard although there is commonality across some in wide use. Metadata systems have also been developed, specifically designed to manage metadata i.e. to provide facilities for input, update, retrieval and reporting of the directory level items. GTOS must have

[1] The term metadatabase is often used more loosely, applying to a catalogue or index e.g. the TEMS metadatabase.

guidelines for directory level metadata that will be required for all GTOS products. These should be developed from existing standards in use. Appendix 4 offers a starting point for this, outlining some examples of metadata and identifying issues to be addressed.

GTOS Data Centres will almost inevitably have metadata specifications and systems in use and these will need to be in consistent with the guidelines. As the network of Centres is established, metadata content between the different institutions will need to be harmonised.

Specific mention should be made of JDIMP activities in this area. The Panel view metadata as key to "registering" G3OS datasets (see 8.2) and, in that context, are developing guidelines for metadata at the directory level. This is building on work done at the Australian Oceanographic Data Centre. One element, common to metadatabases, is the use of theme- or key-words to describe the content of datasets (with the aim of facilitating retrieval). Each observing system would be expected to provide a list of keywords.

As noted above, the requirements for dataset level metadata are dependent upon the particular dataset. GTOS-wide standards for this are impractical. Custodians should be encouraged to use standards or practices common in the particular subject matter field (see section 9).

7.3 Actions

- Identify metadata requirements at the directory level (with JDIMP) and produce guidelines
- Build metadata harmonisation methods (with DCs)
- Identify minimum generic metadata requirements at the dataset level for different data types

8. Data Quality

8.1 Policy

All GTOS datasets will be provided with adequate metadata, enabling users to assess quality in light of the intended use.

8.2 Discussion

Judgements of quality of scientific data are traditionally done through a peer review process. Researchers and users with knowledge in the relevant fields will examine the methods of data collection, analysis techniques used, and the manner in which the results have been interpreted. Detailed documentation of the steps taken and the techniques used to ensure and preserve quality at all stages is required to enable an assessment of quality to be made. Given the number of different types of datasets potentially required, it would

be impossible for GTOS to establish an effective quality assessment process to deliver meaningful results in a useful time frame.

GTOS can however ensure that there is sufficient documentation associated with any data and information to allow an assessment to be made by the user community. This builds on the premise that data quality is best defined as "fitness for use" and should be accurately documented to allow an assessment to be made on that basis i.e. taking into account the use which is to be made of the data. Just as the objective in collecting data may influence the collection method, so the prospective use to be made of data has a bearing on their suitability. Thus a dataset judged to be of acceptable quality for one use may be unacceptable for another.

The requirement for this type of documentation at all stages is an essential element of the end-to-end data management concept. Any "good-quality" dataset must carry such information as part of the metadata associated with it.

Specifically, in establishing a dataset as a GTOS product (see Appendix 1) it should be reviewed to ensure that metadata are complete, i.e. that users may be able to make a quality assessment. Any deficiencies should be identified to the custodian. The proposed operation of the pilot Information Centre (GOSIC) aims to assist this process through carrying out a review as G3OS datasets are "registered " with them. However consideration also needs to be given to ensure that there is a formal recognition of datasets as GTOS products through the Secretariat.

8.3 Actions

- Ensure metadata items contain the information required to allow potential users to judge if the data or information is of acceptable quality.
- Establish GTOS "registration procedures" for datasets and information products.

9. Data Harmonisation

9.1 Policy

GTOS datasets should be harmonised to the extent possible to allow integration of national and regional datasets into a usable global information resource.

9.2 Discussion

Harmonisation seeks to bring together various types, levels and sources of data in such a way that they can be made compatible and comparable, and thus **useful** for decision

making. Harmonisation differs from standardisation in that it does not impose a single methodology or norm, but rather seeks to find ways of integrating or making "an agreeable effect" from information gathered through disparate methodologies.

Harmonisation has been a concern of GTOS since the start - an early meeting of the GTOS Planning Group was hosted by the UNEP Office of Harmonisation of Environmental Measurement in 1994. An excellent summary of the needs and issues in harmonisation was presented at that meeting by E F Roots (1994) and will not be repeated here in detail.

The principal consideration is to find pragmatic ways of making compatible and integrable datasets which have been collected for different purposes under different collection regimes, and using different standards and methodologies. This means avoiding the need to convert all the data to a single standard (impractical), but rather finding ways to make it usable at some higher level of aggregation or generalisation. For some numeric or spatial datasets this could be as simple as applying conversion factors, or a change of map projection. More often in dealing with terrestrial ecosystems data, it is necessary to establish a correspondence between differing classification systems or terminology sets. Common examples are soil, vegetation, land use, and ecological zonation, or descriptive data elements such as threats to habitat, status of species, etc. With biological species data, harmonisation seeks to find a correspondence between similarly described taxa without resolving all taxonomic minutia. Spatial variables which have been collected on different spatial frameworks (e.g. provinces vs. watersheds) may need GIS processing for harmonisation.

A good example at the global level is a handbook of compatible methods of estimating greenhouse gas emissions prepared under the IPCC which defines a number of acceptable methods from which countries may **choose** depending on pre-established national economic and social statistical frameworks and standards.

Such harmonisation it is often considered more difficult to achieve for terrestrial observations because of natural spatial inhomogeneity and discontinuity of the phenomena. The appropriate role for GTOS will be in co-ordinating and facilitating the development of harmonisation approaches and protocols for identified core datasets.

9.3 Actions

- Develop and maintain an inventory of all of the principal international standards organisations and international scientific bodies active in harmonising environmental data relevant to the GTOS scope (particularly terrestrial ecology classification systems, taxonomy, instrumentation and data collection methodologies).
- Identify priority areas where lack of harmonisation is hindering the potential usefulness of GTOS data.
- Facilitate and sponsor international expert meetings to develop harmonisation techniques and correspondence tables in key sectors relevant to GTOS.

10. Archiving

10.1 Policy

All GTOS Data and Information must be securely archived along with all relevant metadata.

10.2 Discussion

The preservation of data and information to enable use over the long-term is intrinsic to the concept of GTOS. Archiving is also an essential element of the end-to-end data management framework and Data Centres on the whole will be expected to have archival procedures in place. One possible approach is for GTOS to verify the adequacy of those procedures as part of the process of designating a Centre to be a GTOS Data Centre. A second alternative is to designate specific DC's as "archive facilities" and ensure that every DC is associated with one such facility to which copies of all material to be archived should be forwarded. This second approach has the advantage that verification and monitoring of procedures is needed at a smaller number of Centres. It also serves to provide the additional security of off-site storage for copies of material.

In the end-to-end process, there are potentially several points at which material should be archived. These points vary depending upon the dataset(s) and processes but should be clearly defined and documented in the overall data management plan in effect. At all stages and in all cases, relevant metadata must be included in the archive material to ensure that the data and information can indeed be used in a meaningful fashion at some later date.

10.3 Actions

- Consider designating a number (at least 2) DC's as GTOS data archive facilities
- Develop guidelines for archival requirements for GTOS products
- Ensure that DC's have archival procedures in place which are consistent with GTOS guidelines

11. GTOS Support.

11.1 Policy

GTOS will provide effective information systems and processes that support and facilitate agreed GTOS functions and activities.

11.2 Discussion

A number of the functions and activities of GTOS outlined in the GTOS Plan, and Sections 3-10 of this Data and Information Plan require or imply the need for information systems support to ensure that the function can operate smoothly and sustainably. Primary amongst these are:

- *Committee Support:* This requires a system to maintain up-to-date terms-of-reference, membership (including contact information) and schedules of all GTOS Panels, Working Groups, Task Groups and the like, as well as associated distribution of documents, agendas, and proceedings.

- *GTOS Data Centres Directory:* As part of the process of designating and linking GTOS Data Centres, a Data Centres Directory must be developed and maintained providing summarised information on the roles, functions, and capacity of each Centre and links to related GTOS datasets and products in the Metadatabase. (Suggested content of the Centres Directory is provided in Appendix 3).

- *Resource Directory:* As noted in Section 9 - Harmonisation, it is important for GTOS to be aware of existing and developing standards related to terrestrial ecosystems information, as well as a range of other scientific issues. A central resource directory is needed to hold institutional and contact information obtained from such action items as in Section 9.3, as well as information and links to institutions having relevant expertise, potential data sources and Data Centres identified by the Science Panels or other bodies.

- *TEMS:* The Terrestrial Ecological Monitoring Sites (TEMS) database is a directory of data (metadatabase) describing monitoring stations and their activities. Proposed actions relating to TEMS are detailed in the GTOS Implementation Plan and are not repeated herein.

The above are all examples of co-ordination and management functions which are required for the operation of GTOS. The question of funding for the Secretariat must be addressed before any long-term strategy can be planned realistically. Thus the actions listed below have been restricted to that pertaining to DCs and a general item concerning systems to provide GTOS support. It should be noted that the Secretariat has already set up some information infrastructure, including a Web page which can be expanded and linked to other sites as requirements are defined.

11.3 Actions

- Design and implement a Data Centres Directory (as Data Centres are designated)
- Plan for systems to support GTOS operations

12. Implementation

12.1 Approach

It is generally recognised that resources for GTOS are scarce. Notwithstanding, it is an urgent imperative to take concrete steps towards becoming an operational observing system complementary to GCOS and GOOS. The development of GT-Net and the demonstration projects to be undertaken by participating networks begin to address this need and GT-Net may be viewed as a prototype for GTOS operation. Clearly, GT-Net activities will address data and information management issues. In fact, some are explicitly identified as key GT-Net activities - e.g. defining a policy on data and information access, sharing and exchanging data, and developing a set of standards for metadata. The timing of the development of this Plan and GT-Net presents an opportunity, over the next 1-2 years, to develop and evolve GTOS policies and procedures with practical experience and to confirm their appropriateness in an operational context.

The principles guiding the implementation of the Plan presented here emphasise **involving** and **evolving**: that is involving the GT-Net developments, existing centres of expertise in observing terrestrial eco-systems, and other interested parties, and evolving the policies and procedures as on-going experience is gained. Thus the policies, guidelines and practices presented here must be seen as preliminary, subject to further development and augmentation (especially of detail).

12.2 Process

As discussed in Section 3, the GTOS Steering Committee, Secretariat and the JDIMP are established and have been addressing data and information management issues at a conceptual level. It is recognised that a GTOS Panel on Data and Information Management is needed for effective implementation and on-going oversight of practical day-to-day policies responding to the needs of GTOS partners. The DIMP is one of several priority Panels identified by the Steering Committee. A first draft of the composition and Terms of Reference for such a Panel is included in Appendix 2.

Given current GTOS constraints, it is probable that a fully resourced Panel may take several years to be realised. In order to move ahead in light of the “evolve” principle and the need to interact with GT-Net as it develops, it is recommended that as an interim measure, a Working Group consisting of SC members should be established to take the lead for all elements of the initial implementation detailed in this Data and Information Plan. The Group would likely invite participation and input from others, for instance, outside experts from the GTOS Roster and managers of GT-Net Data Centres. This composition is similar to what would be expected of a standing Panel (see Appendix 2). The Working Group would operate for 1-2 years with responsibilities (see below) which

would include confirming the composition and terms of reference for its successor Panel, who will then be able to build on and evolve from the initial work.

12.3 Work Program

The overall GTOS Plan catalogues a set of priority actions. The action items listed in the preceding sections of this document are at a more detailed level and are concerned only with the establishment of an infrastructure and system for Data and Information management. These tasks must be undertaken in concert with activities in other areas, recognising that there is overlap. In some cases, considerations of data and information management are the primary focus and driving force whereas in others, they are more subsidiary in nature. The former are dealt with in the first three groups of tasks in the following Table. The specific action items addressed by each task are identified in the second column, along with the section reference. Where draft starting point material has been included in this document, it is indicated in the third column. The fourth task group lists those action items which need co-operative activity.

Task	**Related Action Items**
Task Group 1: To establish a Data and Information Management Panel	
Task 1.1: Define composition, ToRs, mode of operation for SC approval	(sec 3) Create a GTOS Data and Information Management Panel
Task Group 2: To draft the baseline documents regarding GTOS policies, guidelines and procedures to be tested and refined in Task Group 3 below	
Task 2.1: Draft criteria and procedures for designation of GTOS products	(sec 8) Establish GTOS "registration procedures" for datasets and information products
Task 2.2: Draft policy for data access and release	(sec 6) Finalise and publicise GTOS data access and release policy
Task 2.3: Draft criteria and guidelines for designation of GTOS Data Centres	(sec 3) Assess the needs and benefits for designation of different categories of DC (sec 3) Develop criteria for designation of DCs (sec 10) Consider designating a number (at least 2) DC's as GTOS data archive facilities
Task 2.4: Draft guidelines for custodianship agreements	(sec 5) Develop detailed minimum requirements for a GTOS dataset custodian in the form of a proforma custodianship agreement (sec 5) Establish procedures for assigning, managing and reviewing custodians
Task 2.5: Draft guidelines for directory level metadata	(sec 6) Develop guidelines for metadata requirements to enable user browsing and ordering (sec 7) Identify metadata requirements at the directory level (with JDIMP) and produce guidelines (sec 8) Ensure metadata items contain the information required to allow potential users to judge if the data or information is of acceptable quality
Task Group 3: To test each of the above in practice through co-operation with GT-	

Net, TOPC and other GTOS projects as appropriate	
Task 3.1 - 3.5: Refine each of the above in light of pilot application	(sec 6) Build pro-forma agreements to use a starting point for any necessary bi-lateral negotiations (sec 6) Define user needs for mechanisms and tools for browsing and ordering (sec 6) Ensure DCs meet user needs for browsing and ordering (sec 7) Build metadata harmonisation methods (with DCs) (sec 10) Develop guidelines for archival requirements for GTOS products (sec 10) Ensure that DC's have archival procedures in place which are consistent with GTOS guidelines (sec 11) Design and implement a Data Centres Directory (as Data Centres are designated)
Task 3.6: Present documents to Steering Committee for approval	as above
Task Group 4 - Co-operative Tasks	
JDIMP	(sec 3) Continue GTOS representation on JDIMP, at a level to allow full participation
Science Panels	(sec 3) Develop linkages with Science Panels
Panel on Data and Information Release	(sec 4) Ensure analysis of user needs clearly identifies data and information requirements (sec 4) Review outcomes of user needs analyses to identify "core dataset" requirements
International subject matter bodies	(sec 7) Identify minimum generic metadata requirements at the dataset level for different data types (sec 9) Develop and maintain an inventory of all of the principal international standards organisations and international scientific bodies active in harmonising environmental data relevant to the GTOS scope (particularly terrestrial ecology classification systems, taxonomy, instrumentation and data collection methodologies). (sec 9) Identify priority areas where lack of harmonisation is hindering the potential usefulness of GTOS data. (sec 9) Facilitate and sponsor international expert meetings to develop harmonisation techniques and correspondence tables in key sectors relevant to GTOS.

13. Issues and Concerns

Compared to the other Observing Systems, GTOS is relatively immature with regard to defining its scope and modus operandi. The domain of "GTOS Data", the intended user community, and the scope and systematics of observing and assessment methodologies continue to be matters under discussion. Within the agreed general concepts indicated in the GTOS Plan, a number of scientific and methodological issues will likely remain un-resolved for some time, and will evolve as experience is gained and opportunities arise. In some respects, therefore, it is premature to develop data and information policies and plans at a detailed level (for instance, specifications for quality management practices).

In discussions at the Working Group for Data and Information at the Steering Committee meeting in June 1998, in informal feedback on the first draft of this Data and Information Plan, and in earlier fora, a number of information management issues have been raised on which there is not yet a consensus. Some key examples follow:

- Identification, "registration" or "labelling" of GTOS Products, especially regarding whether primary datasets can be considered GTOS products without raising concerns of ownership (See Section 2.2 and Appendix 1).

- Whether there is a need for "types" or classes of designated GTOS Data Centre, and to what degree all DCs should be harmonised with regard to information management.

- Management of non-technical concerns over data "ownership" and appropriate citation of sources, and how to limit the use of GTOS data and products to legitimate scientific and policy development purposes.

- Extent to which GTOS can have a consistent approach with regard to high quality and reliability of data without imposing standards or overly restricting the methodological options of partners.

- Setting priorities for data and products, and identifying target "core datasets".

Another issue, not strictly related to information management is the identification and promotion of the benefits of becoming a GTOS partner and/or Data Centre. At the present time benefits can only be seen as "soft". Easy access to the information holdings of other partners is one such potential benefit that needs to be made more tangible, e.g. through the early development and distribution of a metadatabase.

These issues are flagged to further emphasise the need to involve and evolve, and to be "ruthlessly practical" (Section 2.1) in approaching data and information management in the short term, so as to make progress while these issues are being gradually addressed and resolved.

Appendix I: GTOS Datasets and Products

(This Appendix is intended to provide background for task 2.1)

Given the proposed scope of GTOS activities and the breadth of the targeted user community, the requirements will cover the whole range of what is meant by the terms "data" and "information", from basic observations in primary datasets through processed data in derived datasets to model outputs and indicators. Thus datasets themselves and the results of processing and analysis may all be considered GTOS products.

In the case of derived datasets which will be produced from activities facilitated by GTOS, the situation seems clear. Any such dataset will be a GTOS product. An example is the estimates of Net Primary Productivity from the GT-Net demonstration project. A basic tenet of any GTOS Network should be that any such outputs are designated as GTOS products but ownership will remain with the Network members and will be acknowledged.

The same argument may be applied to new primary datasets which will be compiled through GTOS Network activities (such as those to be generated in the GT-Net demonstration project). These too will be GTOS products. This should work in the same way as in the preceding case i.e. "ownership" will remain with the dataset generators and will be acknowledged.

There is also the question of existing datasets, both primary and derived, which will be used to develop GTOS products where they are identified as required to meet a priority need. Datasets currently identified by TOPC are examples of this. A formal approach should be made to the dataset owners asking that the dataset be made accessible to GTOS. Such datasets may not be GTOS products in the strict sense of "produced by GTOS" but if they are used by GTOS they are GTOS datasets in the sense that they have received some form of approval. The suggestion has been made that, in acknowledging ownership of any such dataset, it should be flagged as "a contribution to GTOS".

Thus for all of the above, mechanisms are needed to ensure that, where appropriate, a GTOS designation is attached to any product (including datasets) without usurping ownership in any way. Any such dataset would need to be managed according to GTOS policies, including having the required metadata at the directory level.

Appendix II: GTOS Data and Information Management Panel

(This Appendix is intended to provide a starting point for task 1.1)

Composition of Panel

The Data and Information Management Panel should be composed of a core of Steering Committee members with expertise and experience in information management issues, and representatives from the designated GTOS Data Centres. The Panel can further be augmented by outside experts chosen from, for example, the GTOS Roster of Experts, from GCOS and GOOS, and from other bodies with common interests such as IGBP-DIS and WGISS.

To be effective the size of the Panel should strike a balance - large enough to have the range of expertise needed to span the likely issues, but not so large as to be unwieldy. A total of about 12 people might be considered with at least 4 coming from the Steering Committee.

Membership should rotate on a regular basis.

Terms of Reference

The following should be included in the terms of reference for the Panel.

- Responsibility for the current Work Programme items related to Data and Information.
- Review and evolution of the GTOS Data and Information Plan - for approval of the Steering Committee and incorporation into the overall GTOS Implementation Plan.
- Interaction with the Science (and other) Panels as required, for example to identify priority user data and information needs.
- Representation of GTOS on the Joint Data and Information Panel - through designation of one or more Steering Committee members of the GTOS Panel.
- Advising the Steering Committee on emerging Data and Information issues

Mode of Operation

Recognising the limited resources of GTOS, the Panel will be expected to conduct its work as much as possible by email communication and “electronic meetings”. In person meetings would be most efficient if held in conjunction with other related events, such as Steering Committee meetings.

The Panel should be authorised to commission small task groups to address selected priority data and information issues, and prepare discussion papers or recommendations for the Panel. Task Group members need not be Panel Members, and Task “Groups” could consist of a single individual if appropriate.

Appendix III: GTOS Data Centres

(This Appendix is intended to provide a starting point for task 2.3)

Designation of Data Centres

In designating a GTOS Data Centre (DC) consideration should be given to defining both its **data domain** and its **processing domain**.

The **data domain** identifies the range of data categories with which the Centre deals whether it be for primary data collection and assembly including quality assurance, and/or the use of secondary data, integration and product production. This domain may also be circumscribed geographically - e.g. "wildlife data in protected areas of South East Asia".

The **processing domain** defines the principal data management functions and capacity of the Centre, highlighting those in which the Centre has particular expertise. These processing functions or operations might include:

Assembly - building up and organising a dataset from primary observations for a single data type

Integration - bringing together data from multiple data sources into a compatible database

Analysis - selection, processing, computation and other manipulation of the data - usually to produce an information product. Analysis functions include:

- error detection and correction
- summarisation and generalisation
- statistical analysis
- time series and trend analysis
- spatial analysis (and GIS)
- image analysis (e.g. for remote sensing)
- mathematical modelling
- visualisation

Product generation - integration and analysis to produce a new data set, or other product

Data and product distribution - including electronic access as well as digital media distribution

Archiving - preserving, and providing access to, valuable datasets and products.

Checklist of Criteria for Agreements/Designation of Data Centres

Manages designated GTOS datasets and products

Conforms to GTOS data access/release policies

Follows GTOS data management polices (e.g. with respect to quality assurance and archiving)

Archives GTOS data sets (or provides facilities to archive)

Has analysis /processing capacity appropriate to the needs of GTOS (which can be accessed or utilised by GTOS partners under reasonable conditions)

Possible Data Centre Categories

(The category designation should be based on the primary or major functions of the Centre)

Primary Centre: provides all aspects of end-to-end data management for a major data domain of value to GTOS.

Specialised Centre: engages in management of a select, but key, GTOS dataset

Archival Centre: specialises in providing archive facilities and access to archival datasets.

Analysis Centre: provides functions or facilities for data integration and analysis, e.g. global modelling, trend analysis.

Affiliated Centre: co-operates with GTOS in useful ways, but does not (for technical or administrative reasons) meet all of the DIM policies

Contents of a GTOS Data Centre Directory

Each GTOS Data Centre should be fully described in a Centres Directory (or institutional meta-database). Minimum contents of such a Directory are as follows:

Name of Centre: (full name, as well as short name or abbreviation/acronym)

Responsible Authority: (organisation responsible for the Centre - e.g. Ministry, University, etc.)

Responsible Contacts: (including full address, email etc.)

Mandate: (description of objectives, purpose of the organisation)

Data Domain: (GTOS datasets maintained, geographic limits, etc. - linked to GTOS directory level metadatabase)

Data Access: (policies and conditions of access, directory level metadatabase used)

Functional Domain: (available processing functions, analysis capacity, conditions of access)

Expertise: (description of specialised expertise (human resources) available for consultation)

Appendix IV: Directory-Level Metadata

(This Appendix is intended to provide background material for task 2.5)

1 Introduction

It is commonly agreed that in this "information age" data itself is a resource and an abundant resource at that. As with many natural resources, it is important for a potential user to know first of the existence of data, and also to know where the data are to be found, their condition, and other information to determine whether or not they can be used for the task at hand. This knowledge about existing data and how it might be accessed and used, has become more and more important as funding for systematic data collection activities has been reduced in many jurisdictions.

Metadata are "data about data", describing such things as the location, sources, content, quality, condition of existing data. Metadatabase systems are systems specifically designed to manage metadata i.e. to provide facilities for input, update, retrieval and reporting of data about data. Such systems are used at a variety of levels, for example, within a single institution or organisation to organise and maintain their own data holdings in order to protect and maximise the investment in organising and structuring data. They are also used on a broader level to provide a mechanism through which data producers can ensure that potential users can be made aware of existing data and how it might be obtained. The systems may also vary in the types of data which are described, for example, books, reports, maps, digital files, etc. Thus there is a wide range of relevant work, from bibliographic systems to handling of digital imagery.

2 Sample Metadata Developments

2.1 NASA Global Change Master Directory

The Global Change Master Directory (GCMD) of NASA is a comprehensive directory of material relevant to global change research. It uses a Directory Interchange Format (DIF) which has been widely used for exchange of metadata. The content of the DIF is organised into a nested structure using prescribed “labels” to identify the beginning and end of information fields. The term “DIF” is also used to refer to an entry in the directory i.e. the description of a dataset is referred to as a DIF. The standard GCMD DIF allows for some 43 fields (many of which are fairly specific to remote sensing data) which can be grouped roughly as follows:

identification: (addresses, contact names and other information about the agencies and people responsible for the dataset)

spatial reference: (geographic scope and location of the data, map projections used and the like)

distribution: (information on conditions and methods of access to the data)

metadata reference: (information on data content, resolution, scientific purpose, etc.)

As can be seen from the examples given below, these are the types of information common to many metadatabase developments though the actual fields will vary.

A limited number of DIF fields are considered mandatory, and many permit narrative descriptions or comments. Some fields have controlled vocabulary - e.g. prescribed allowable keyword lists. Again, both these features are common to other metadatabase developments.

2.2 United Nations Environment Program

Within UNEP, a metadatabase was developed as part of the GRID programme. A "dataset" is defined as a collection of data and accompanying documentation maintained at a single source, where a collection of data refers to a minimum of one or a series (no maximum) of “data members” which relate to a specific theme or geographic region in terms of physical area covered. Information held about a dataset includes:

identification: (name of dataset, and institution holding it)

contact references: (names, details of address of contact, access conditions)

geographic coverage: (general location , e.g. continent or country, and /or latitude-longitude bounding rectangle)

general description: (subject keywords and free text summary of dataset contents)

A "member" is regarded as equivalent to a data file, a paper report, a map or other unit of data, and is always a component of a particular dataset. Thus, members are the lowest-level, "concrete" data entities that could actually be requested by a potential user. Information held about a member will vary with the type of member, for instance, metadata items for a raster data file will include resolution and number of rows and columns, for a vector data file they will include geo-referencing details, and so on.

A separate section of the metadatabase is maintained containing information on “institutions”, that is, the centres that hold data. This includes information on the overall scope and nature of scientific programmes and the information management capacity of the centre. This could provide a useful model for a GTOS Data Centres Directory (see Section 11).

2.3 Federal Geographic Data Committee (FGDC)

In the USA, the FGDC developed a standard for digital spatial metadata through a consultative process over a 2-year period starting in 1992 and it is now mandatory for all US Federal agencies. The standard provides a common set of terminology and definitions for the documentation of the data. It establishes the names and definitions of data elements and groups of data elements to be used for these purposes, and information about the allowable values for the data elements. The standard also defines which data elements are mandatory, mandatory under certain conditions, and optional (i.e. included at the discretion of the data provider).

The standard is quite extensive specifying the structure and expected content of some 220 data elements. These can be roughly grouped as follows:

Identification: (basic information about the data set, including the dataset name, geographic area covered, currency, and rules for obtaining or using the data)

Data Quality: (information which assists in assessing the usefulness of the information for the users purpose, including, the positional and attribute accuracy, completeness, consistency, information sources and methods used to process the data)

Spatial Data Organisation: (spatial representation methods, e.g. the method used to represent spatial positions directly (such as raster or vector) and indirectly (such as street addresses or county codes) , the number of spatial objects in the data set and so on)

Spatial Reference: (description of the reference frame for, and means of encoding, co-ordinates in the data set including map projections parameters, grid co-ordinate systems and resolution, and the horizontal and vertical datum)

Entity and Attribute Information: (information about the content of the data set, including the entities types and their attributes and the allowable attribute value domains)

Distribution: (information about obtaining the data set, including contact addresses, available formats and media, online access, and fees for the data)

Metadata Reference: (information on the source of the metadata entry and its most recent up-date)

There is a systematic mapping of the fields of this to those of the GCMD DIF.

Details of information to be reported and tasks to be performed are in the Spatial Data Transfer Standard (Federal Information Processing Standard 173).

2.4 The Australia New Zealand Land Information Council (ANZLIC)

ANZLIC has developed a definition of the appropriate elements for a national land and geographic data directory system through a consultative process which began in 1995. The approach is deliberately less ambitious than that of the FGDC described above but is, as far as possible, consistent with the FGDC guidelines.

The core elements of the definition are grouped into 9 categories:

Dataset: (title, custodian, jurisdiction)

Description: (abstract, search word(s), geographic extent with name(s), or geographic extent defined by polygon(s))

Data Currency: (beginning date, ending date)

Dataset Status: (progress, maintenance, update frequency)

Access: (stored data format, available format type(s), access constraints)

Data Quality: (lineage, positional accuracy, attribute accuracy, logical consistency, completeness)

Contact Information: (name, title, organisation, address, contact numbers)

Metadata Date: (date metadata was prepared or updated)

Additional Metadata: (optional information)

Note that the JDIMP metadata pilot project is being led through the Australian Oceanographic Data Centre and uses the above categories as a starting point.

3 Issues

3.1 Level of agreement on metadata content

The above has given some samples of activities in construction of metadatabases. It is an area in which there has been increasing interest in the recent past and there are a number of data standards emerging, but no single one is adopted generally. As shown in the preceding paragraphs, there are several well-documented examples which have common elements.

There appears to be general consensus that the way forward is to have some form of metadata standards harmonisation which will at least facilitate movement between metadatabases. This is what GTOS will have to put in place to facilitate information

exchange and access across Data Centres. The important principle is to include the information needed to determine if the dataset is **potentially** useful for the user's purposes.

3.2 Standard Terminology and Keywords

The use of clearly stated and well-defined terminology in constructing a metadatabase is essential. Open-ended text searching to determine dataset content is a very hit-and-miss method. For environmental data, existing keyword lists serve reasonably well for the broad category levels but frequently present problems as the need to add detail arises. GTOS may need to extend or modify existing vocabularies established or underdevelopment, e.g. by the EEA, CEISIN or UNEP. Any such lists should also be open-ended i.e. allow for additions and evolution, but existing metadata must be kept consistent with any changes made.

3.3 Geographic Location

An important item of metadata describes the geographic area covered. The use of geographic names for this presents problems in the standardisation of the names of countries and regions, and also in the way in which the metadata has to be handled. For example, an metadata entry which describes an entry as relating to "East Africa" could be of interest to someone searching for information about "Africa", or for information about "Kenya". Consideration should be given to using an existing standard for country names and abbreviations such as from ISO or the numeric codes used by FAO. This can allow for consistency in searching.

Another commonly used method is the specification of a "bounding rectangle" or point location using, for example, latitude and longitude.

3.4 Metadata Output and Exchange

The Directory Interchange Format mentioned above defines a formal syntax that helps to ensure that the metadata is as complete and unambiguous as possible and this syntax is generally being adopted for exchange of metadata at all levels. The output/exchange files are in ASCII format consisting, for each dataset, of each attribute or field name followed by the value, with a colon separating the two. Although it is unattractive as printed output, it is possible to read it electronically without ambiguity. This format can be considered a de facto exchange standard, and can be employed to easily import existing metadatabases from participating DCs.

3.5 Availability of Tools

A variety of software packages have been developed and used for the management of metadatabases. Currently in countries where there is readily available access to the Internet, there is substantial development activity in providing metadatabase access tools using those protocols. These developments tend to concentrate on facilitation of metadata

entry and development of search interfaces (including implementation of spatial queries) to existing metadatabases.

3.6 Level of Effort

The construction and maintenance of a metadatabase involves significant effort often seen as unglamorous extra work beyond the initial data assembly. It is important that the burden of supplying metadata not be too onerous. Experience has shown that the greater the number of mandatory fields to be entered, the greater the likelihood is that no entry will be made at all. The rather simplified ANZLIC structure reflects this thinking. A balance must be established between the need to have sufficient information for a user to identify potentially valuable datasets, and the need for ease in entering and maintaining the metadata. In this regard it is important to make the distinction between the “directory level” metadata - referred to here, which must be kept simple, and the “dataset level” metadata (sometimes called co-data) which must be available with the dataset. This dataset level metadata must contain all the information needed to correctly use the dataset - calibrations, off-sets, origins, assumptions, classification systems, terminology sets, taxonomies, geo-referencing schemes, instrument parameters, legends and coding schemes, etc. etc., as appropriate to the dataset.

Appendix V: List of Acronyms

DC	Data Centre
DIMP	Data and Information Management Panel
FAO	Food and Agriculture Organization of the United Nations
GCOS	Global Climate Observing System
GOOS	Global Ocean Observing System
GOSIC	Global Observing Systems Information Center
GTOS	Global Terrestrial Observing System
GTOSSC	GTOS Steering Committee
G3OS	The three Global Observing Systems (GCOS, GOOS and GTOS)
ICSU	International Council of Scientific Unions
IGDB-DIS	International Geosphere-Biosphere Programme - Data and Information System
IPCC	Intergovernmental Panel on Climate Change
JDIMP	Joint Data and Information Panel
JSTC	Joint Scientific and Technical Committee
TEMS	Terrestrial Ecological Monitoring Sites
TOPC	Terrestrial Observation Panel for Climate
UNESCO	United Nations Educational, Scientific and Cultural Organisation
UNEP	United Nations Environment Programme
WGISS	Working Group on Information Systems and Services
WMO	World Meteorological Organisation