

A proposal for Data Management in GEOTRACES
Report from Liverpool Meeting Dec 2005

Contents

1. Executive summary
 2. Preamble
 3. Objectives of the Data Management Meeting
 4. GEOTRACES overview
 - 4.1 GEOTRACES goals and themes
 - 4.2 GEOTRACES data types and their requirements
 5. Data management for GEOTRACES
 - 5.1 Introduction
 - 5.2 Oversight and organizational requirements
 - 5.3 Data Flow
 - 5.4 Data sharing policies
 - 5.5 Data centres
 - 5.6 Special requirements of process studies
 - 5.7 Timing and cost
 6. Recommendations
- Appendix 1 list of attendees
- Appendix 2 detailed list of data types
- Appendix 3 individual presentations
 - A3.1 GEOTRACES scientific overview – Chris Measures
 - A3.2 Hydrographic and nutrient data – Jim Swift
 - A3.3 Data management overview – Roy Lowry
 - A3.4 Data Centre interest in GEOTRACES
- Appendix 4 Abbreviations and definitions

A proposal for Data Management in GEOTRACES

Report from Liverpool Meeting Dec 2005

1. Executive summary

Recognizing the importance of data management for the GEOTRACES project, the GEOTRACES Planning Committee convened as one of its first activities a meeting designed to launch GEOTRACES data management. Initiation of data management is a crucial prerequisite for GEOTRACES field research. The meeting described in this report resulted in recommendations for a GEOTRACES data management system and data policies to guide GEOTRACES scientists. The data management system should include a Data Management Committee, a Data Liaison Officer in the GEOTRACES International Project Office, Data Specialists on GEOTRACES cruises and associated with process studies, and two Data Assembly Centres that will manage GEOTRACES data. Proposals are also put forward for metadata and data collection; time scales for data and metadata delivery to data centres and participants; timescales for public release of data; and a data sharing policy. Appendices set out much of the background thinking on which these proposals are based. This report will be delivered to the GEOTRACES Scientific Steering Committee for consideration and action at their first meeting, in mid-2006.

2. Preamble

At the meeting of the GEOTRACES Planning Committee held in Vienna in May 2005, Chris Measures and Raymond Pollard volunteered to convene a group of experts to discuss the data management requirements for GEOTRACES, prior to setting up a formal Data Management Committee (DMC). The resulting data management meeting was hosted by the British Oceanographic Data Centre (BODC) in Liverpool, UK on 30 Nov – 2 Dec 2005 and this report is the output of that meeting. A list of attendees is given in Appendix 1. A list of abbreviations and brief explanations of technical data management jargon are given in Appendix 4.

In order to propose a data management system it is first necessary to have a good idea of the data types, their characteristics and quantities that must be handled. The meeting therefore began with an overview of GEOTRACES and then listed in some detail the expected data types and their collection techniques. The details of data management—from collection to final archiving—were then discussed. This report follows the same structure. Much of the detailed material and presentations will be found in appendices, so that the main conclusions and data policy proposals are relatively short. A significant number of the recommendations have been lifted, with little or no change, from previous data management meetings or programmes (see www.jhu.edu/scor/DataMgmt.htm). Meeting participants recommended adopting, as much as possible, successful approaches used by other projects.

3. Objectives of the Data Management Meeting

- Detail the data types and data management requirements of GEOTRACES

- Review the experiences of previous projects and how these can contribute to forming a GEOTRACES data management system
- Specify data management policies for GEOTRACES
- Design a GEOTRACES data management system, or set out and compare alternatives
- Document the next steps and the time scale on which progress is desirable
- Create a report for consideration by the GEOTRACES Scientific Steering Committee.

4. GEOTRACES overview

4.1 GEOTRACES goals and themes

The mission of the GEOTRACES programme is **to identify processes and quantify fluxes that control the distributions of key trace elements and isotopes in the ocean, and to establish the sensitivity of these distributions to changing environmental conditions.** GEOTRACES has three primary goals

- To determine global ocean distributions of selected trace elements and isotopes, including their concentration, chemical speciation, and physical form, and to evaluate the sources, sinks, and internal cycling of these species to characterize more completely the physical, chemical and biological processes regulating their distributions.
- To understand the processes involved in oceanic trace-element cycles sufficiently well that the response of these cycles to global change can be predicted, and their impact on the carbon cycle and climate understood.
- To understand the processes that control the concentrations of geochemical species used for proxies of the past environment, both in the water column and in the substrates that reflect the water column.

These goals will be pursued through complementary research strategies, including observations, experiments and modelling, organised under three themes:

Theme 1—Fluxes and processes at ocean interfaces - Atmospheric deposition; continental run-off; the sediment-water boundary; ocean crust

Theme 2—Internal cycling - Uptake and removal from surface waters; uptake and regeneration in the sub-surface ocean; regeneration at the seafloor; physical circulation

Theme 3—Development of proxies for past change - Factors controlling “direct” proxy distribution in the ocean; factors influencing the distribution of “indirect” proxies in the ocean; paleoceanographic tracers based on sediment flux.

4.2 GEOTRACES data types and their requirements

GEOTRACES will be global in scope, consisting of **ocean sections** complemented by **regional process studies**. Sections and process studies will combine fieldwork, laboratory experiments and modelling. Sections will be planned to cross regions of prominent sources and sinks (such as dust plumes, major rivers, hydrothermal plumes and

continental margins), to sample principal end-member water masses, and to enter the major biogeographic provinces.

Data requirements for ocean sections will be relatively standard and well defined, consisting of vertical water sampling with stations spaced several degrees apart, with additional data fields from on-board continuous sensors. There is much experience of handling these data types. In particular, the CLIVAR and Carbon Hydrographic Data Office (CCHDO) at the Scripps Institution of Oceanography (SIO), originally the WOCE Hydrographic Data Assembly Centre (DAC), has considerable experience of handling vertical water profile data. GEOTRACES should make use of the CCHDO, which is already funded to manage profile data internationally.

Process studies will be more diverse than ocean sections. Some studies will be similar to but shorter than ocean sections (in time and length), whereas others will involve measurements and observations that will yield more varied types of data requiring specialised handling. Process studies will involve a large number of radically different data types, for example, video and still cameras, pore water samplers, cores, sediment traps, mooring systems, etc. Process studies will probably be co-operative ventures involving several countries with varying capabilities and resources. In particular, these studies may be undertaken by groups that lack experience of data management requirements for international projects. Some studies will be subject to Exclusive Economic Zone (EEZ) data-release restrictions. For all these reasons, managing the data sets produced from the process studies is expected to be significantly more complex than for section data. Nevertheless, the ocean data management community has experience in handling all these data types, and indeed how to allow for the incorporation of data types not envisaged when the project was set up (“data orphans”). The meeting heard presentations from the BODC, PROOF, CCHDO and the CIESEN about their data management capabilities (Appendix 3).

To fully realise the goals of GEOTRACES it is important that all data be made available as soon as possible to shipboard personnel and shore-based researchers as well as to the general scientific community. Lists of data types are given in Appendix 2, and proposals to manage the data are developed below.

5. Data management for GEOTRACES

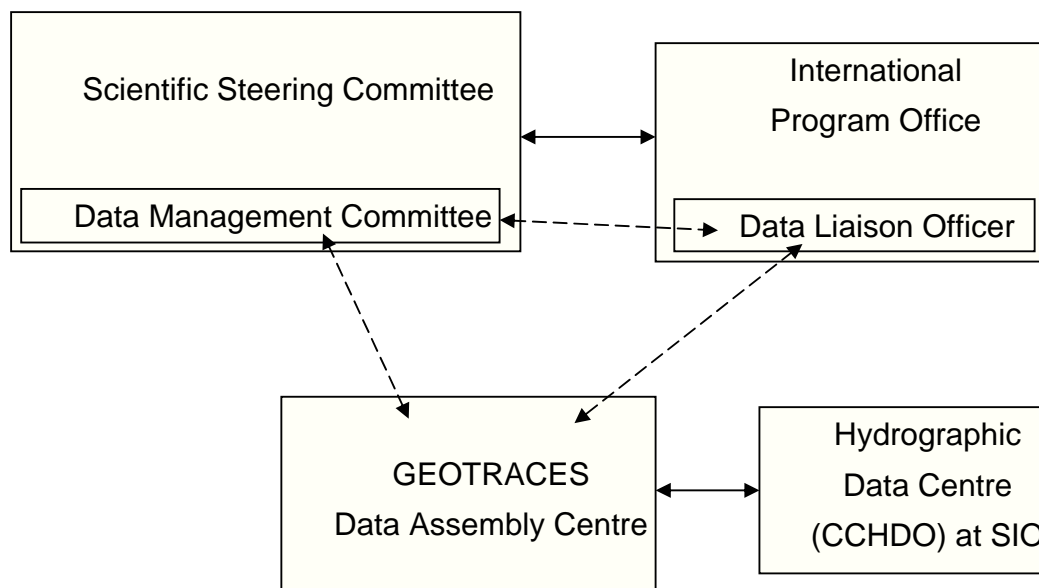
5.1 Introduction

A number of presentations made during the meeting are included or summarized in Appendix 3. Much reference was also made to lessons learnt from previous projects, which are well documented at www.jhu.edu/scor/DataMgmt.htm. Here we present for the GEOTRACES SSC our proposals on Data Policy and Procedures for GEOTRACES. Further recommendations on matters of detail will be included in due course in a data manual for use by GEOTRACES PIs and for cruise and process study lead scientists.

We shall distinguish between “metadata,” which describe a data set, from the “data” themselves. Metadata comprise the essential information about how, what, where, when and by whom data were produced. Without them, the actual data are worthless.

5.2 Oversight and organizational requirements

The proposed GEOTRACES Data Management System should include a Data Management Committee, a Data Liaison Officer in the GEOTRACES International Project Office (IPO), and two international Data Assembly Centres (DACs), as shown in the diagram. There will be, in addition, national data management structures as nations require. Data Specialists should support cruises and process studies. These elements are now described.



GEOTRACES should establish a Data Management Committee (DMC) comprised of data originators (i.e., observational scientists), data managers (national and international) and data users (including modellers).

The GEOTRACES DMC should have three areas of responsibility (taken from <http://www.jhu.edu/scor/DMReport.pdf>):

- (1) ensure that data are available for project scientific purposes and that data management meets the present scientific needs of the project without compromising future needs
- (2) oversee the compilation of data from individual principal investigators (PIs) and national projects into a long-term, integrated data set that is submitted to an appropriate data archive and may be published in a suitable format (CDROM or DVD are current possibilities)
- (3) address the involvement in project data exchange activities of scientists without access to effective data management infrastructure.

The following draft terms of reference for the DMC are derived from these areas of responsibility, augmented by items from other major projects.

Draft Terms of Reference for GEOTRACES Data Management Committee (DMC)

- Oversee the work of the GEOTRACES Data Assembly Centres (DACs) and the Data Liaison Officer in the GEOTRACES International Project Office.

- Ensure that GEOTRACES creates and maintains an integrated, international data and cruise inventory
- Oversee the compilation of data from individual principal investigators (PIs) and national projects into a long-term, integrated data set that is submitted to an appropriate data archive and may be published in a suitable format (CDROM or DVD are current possibilities)
- Ensure that GEOTRACES data are available for project scientific purposes and that data management meets the present scientific needs of the project without compromising future needs
- Monitor international acceptance of, compliance with, and adoption of, GEOTRACES data policies
- Address the involvement in project data exchange activities of scientists without access to effective data management infrastructure
- Report regularly to and advise the GEOTRACES Scientific Steering Committee (SSC)

Role of IPO in Data Management

We agree with the recommendation of the SCOR/IGBP meeting on data management (<http://www.jhu.edu/scor/DMReport.pdf>) that management of metadata is best handled by a person working at the IPO. This may not be a full-time job, but the time commitment should not be underestimated. A Data Liaison Officer (DLO) should be employed at the IPO from an early stage, with the following responsibilities

- Maintain a list of GEOTRACES cruises
- Keep track of project metadata
- Maintain a catalogue of actual and expected data sets by producing Directory Interchange Format (DIF) or equivalent discovery metadata records (conforming to the ISO19115 standard for metadata)
- Ensure that standardized parameter descriptions are adopted, for example, the BODC parameter usage vocabulary.
- Define dataset boundaries in consultation with the SSC
- Ensure that protocols for naming of cruise, station positions, etc., adhere to a rule system developed by the DMC
- Interact with DAC(s) to coordinate their activities and interactions with PIs
- In particular, ensure timely delivery of metadata and actual data to the DAC(s)
- Contribute to (and possibly maintain) the project web-site

The DLO will be an *ex officio* member of the DMC, and will report to the Director of the IPO and to the DMC.

5.3 Data Flow

Managing data involves a complex set of operations, including collection, calibration, documentation, submission to data centres, quality control, dissemination and archival. Scientists are not usually very good at managing this entire set of important operations

(apart from their prime responsibility to collect top-quality data), and it is strongly recommended that data management professionals be involved in all GEOTRACES data activities from the start. This is known as co-operative data management (sometimes termed “end-to-end” data management, which is the phrase used elsewhere in this document) and is already practiced by, for example, BODC and the Marine Aspects of Global Change (IMAGES) project. In essence, co-operative data management means involving a data centre from the planning stage, including participation on the cruises and/or involvement in data collection, as we now describe.

Responsibility for **data collection** is spread among many individuals – PIs, their technicians, students, etc. But on a cruise it is the Principal Scientist’s responsibility to ensure that metadata and data documentation are completed and delivered to the IPO. This is a major task, and it is strongly recommended that at least one berth on every GEOTRACES section cruise be allocated to an individual with data management experience to serve as a Shipboard Data Specialist (SDS). Similarly, each process study will have a lead scientist, who should be supported by a Process Study Data Specialist (PSDS). Allocation of funding to Data Specialists will pay off amply in the completeness and quality of the GEOTRACES final data sets.

Data Specialists

The Data Specialists should be tasked to help the Principal Scientist (of a cruise or process study) with data issues, and would be briefed before the start of the cruise or project by the Data Liaison Officer. Responsibilities of the Data Specialists would include

- Ensure that suitable log sheets have been provided for all activities
- Assist and support scientists in preparation of metadata
- Maintain regular checks that all logs are being correctly completed
- Assemble all metadata from a section or process study
- Assist with preparation of data files, ensuring that all necessary parameters are included
- Evaluate the quality of data, either by personal expertise or by discussion with PIs, and help to document quality and missing or suspect data
- Facilitate assembly of shipboard data sets and data integrity checks.

This list is far from complete, and needs to be expanded by the Data Liaison Officer, working closely with the DMC. The duties of the Data Specialists will vary from cruise to cruise (and process study to process study) depending on their personal expertise and the details of each activity, but some items will always be included, primarily concerning collection of complete metadata and calibration of the hydrographic data. Data Specialists may be employed by the Principal Scientist or another cruise scientist as part of a grant, employed by the ship-operating institutions or employed by one of the Data Assembly Centres.

Metadata capture

Endorsement of a scientific activity by GEOTRACES requires metadata to be submitted on the shortest possible time scales. Discovery metadata (what was collected where,

when and by whom) should be submitted by project scientists with the help of the Data Specialist to the Data Liaison Officer at the IPO as the cruise is planned and when the research vessel docks at the end of a cruise. Failure to do so should be considered reason to remove GEOTRACES endorsement, as the lack of access to metadata compromises the ability of GEOTRACES to fulfil its goals. Involvement in GEOTRACES should also entail a credible commitment to the timely submission of data to a project-approved database to ensure long-term archiving of the data.

The unique key for a bottle water sample will be a combination of cruise, station, cast and bottle number, where bottle number is the number engraved on or firmly assigned to a particular bottle. This is a valuable reference if a bottle leaks or is otherwise contaminated, particularly for tracer work. The position on the rosette frame should also be recorded, and is usually the order in which bottles were fired. Examples of this and many other detailed requirements are given in Appendix A3.2.

The Directory Interchange Format (DIF) developed by the Global Change Master Directory (GCMD) is a suitable standard for cataloguing datasets and has established storage and query infrastructures. GEOTRACES should adopt this or a similar discovery metadata standard, for example, ISO19115 when a suitable marine profile is available.

Timescales for data submission

The IPO should take the lead role in maintaining a catalogue of project metadata. Operating a project metadata catalogue should be considered a core activity of the IPO and will be a major duty of the Data Liaison Officer. The rationale is as follows.

Both metadata and data need to be submitted to the Data Liaison Officer and the relevant DAC as soon as they are created; metadata about cruises (when, where, who, what will be measured) should be submitted when a proposal is funded and cruise metadata should be submitted immediately after the end of the cruise. The most important reason for this is data security - with the best will in the world individual PIs may lose metadata or even the data themselves and duplication is extremely important to provide backup. Metadata comprise, in part, the detailed lists being made during the cruise or process study and the Data Specialist (working with the lead scientist) should ensure that these lists are printed, copied or scanned on board the ship (or during the process study) and that duplicates are delivered to the DLO immediately.

Cruise reports should be submitted to the DLO within 6 months of the end of the cruise. The cruise reports will be publicly available, distributed through the GEOTRACES portal and, if possible, be assigned a reference digital object identifier (DOI).

The data assembled on board ships should be delivered to the appropriate DAC within one month of the end of the cruise (for exceptions see below). These may be preliminary or partially calibrated data, as accompanying metadata will show, and the most important reasons for such rapid delivery is to start the dialogue with the DAC on the quality and completeness of the data as well as for data duplication and security. Participation in a cruise implies willingness to share data with other cruise participants. The DAC will be able to check completeness with the relevant PI and may be able to help with error checking. It is recognised that, for many data, "one month" delivery is impossible, and a table of expected delivery times will be developed to allow for necessary delays for particular parameters, for example, isotope ingrowth.

It is expected that cruise participants will have access to routine hydrographic parameters as soon as they are available, either shipboard or as soon as available at the DAC. Whether this access extends to the individual TEI data sets of other PIs is an issue that needs discussion and resolution by the GEOTRACES SSC - there are arguments for and against this approach.

The major incentive for rapid data submission is the extra support that will then be available to quality control and to interpret the originator's data. The data will be integrated with the authoritative metadata without personal effort. The DAC will quality control data and may help the originator in detecting problems. Comparison with the data sets of other PIs is one sure way to tease out errors and inconsistencies. Access to related data will aid in linking of individual data sets and promoting scientific collaboration. Data will be professionally maintained, safeguarded and archived.

Timescales for data submission - summary:

Metadata - as soon as created from the planning stage onwards

Data (not finalized) - within 1 month (of collection, or end of cruise), with possible approved extension as tabulated

Cruise report - within 6 months of the end of the cruise

Final detailed data report for each process study (see section 5.6) - within 6 months of the end of the process study

Final data - within 2 years, exceptions possible from the GEOTRACES SSC

Timescales for data release

GEOTRACES must adopt the ICSU principle of free and open data exchange, which will help GEOTRACES achieve its goals. Data release can be split into two categories: (1) release to other participants in the cruise and (2) public release. The DAC will have to hold a list of "participants" (which may include closely associated persons who were not actually at sea, as agreed with the Principal Scientist) for each cruise. In general, data should be available to participants without delay, but the DAC will need to consult the relevant PI or keep blanket permission information with the metadata. Procedures for restricted access to data are well established (e.g., using paired passwords and approved user lists) and will be implemented by the DAC(s). Participants in other GEOTRACES cruises should contact the relevant PI directly to obtain desired data; they will be aware of the existence of the data through the metadata maintained by the IPO. Public release will normally be two years from the end of the cruise or field activity, but should be extended when analytical procedures have inherent built-in delays.

Timescales for data release - summary:

Participants in a particular cruise - as soon as available at the DAC with knowledge and permission of the relevant PI

Public release - within two years of end of cruise (+ extra time for particular data type as approved by SSC)

Validation and quality control

Individual PIs are responsible for quality control of their data and should provide in their metadata notes about any doubtful data values. Questionable data should be flagged rather than discarded. Participants using the data should report questionable data to the DACs (who also will be noting problems), or possibly to the PI copied to the DAC. The DACs will pass data quality problems back to the PIs. The cruise or process study Data Specialist and the Data Liaison Officer should be copied into such correspondence and may be able to help, at least in the documentation.

Experience shows that data quality problems are often revealed once the data begin to be used scientifically, often by individuals other than the data collector. Comparison with other parameters or comparison between data sets at the same location can reveal errors or offsets. This is one good reason for making data available to other participants without delay. More formally, groups of PIs expert in a particular parameter (for example Fe) will be encouraged to apply further quality controls, such as cruise intercalibration.

Archiving

The data from GEOTRACES must ultimately be preserved for posterity. In part this can be done by providing data sets in multiple copies (such as DVDs), but all media degrade on some time scale and the formats and technical specifications of storage devices are constantly evolving. The World Data Centres (WDCs) have the resources to preserve data for the long term and periodically to update the media on which data are held. Meeting participants believe that the World Data Centre for Oceanography, Silver Spring would be the most appropriate data centre to hold most GEOTRACES data. Personnel from this WDC should be encouraged to interact with the DMC from an early time to facilitate eventual long-term data archiving at the completion of the programme.

5.4 Data sharing policies

In light of the above discussion, the formal data sharing policy is as follows:

There is a fundamental trade-off in GEOTRACES - on the one hand, protection of the intellectual effort and time of originating investigators (those who plan an experiment, collect, calibrate, and process a data set to answer some questions about the ocean), and on the other hand, the need to compare various data sets and data types to check their consistency, to better understand the ocean processes involved, and to see how well the numerical models describe the real ocean. The policy adopted by GEOTRACES is a trade-off between these conflicting needs.

GEOTRACES activities require all participating scientists to submit their metadata to the Data Liaison Officer at the IPO as soon as they are available (including cruise/process study details when the proposal is accepted for funding). Any metadata and data produced during the cruise/process study should be made available to participating scientists immediately in preliminary form during the cruise/process study. "Routine data," by which we mean the basic hydrographic parameters, will be made public a short time after each cruise/process study is completed. Preliminary data collected as part of GEOTRACES are to be submitted to the DAC within a month of their collection for the purposes of quality control and data synthesis during the 2-year "publication rights period." Any data collected as part of GEOTRACES should be made publicly available

no later than 2 years from collection, with an extension of this period as specifically granted by the GEOTRACES Scientific Steering Committee (SSC) for particular parameters that require extensive processing after the cruise or process study is completed, and recognising that some nations do not permit release of data collected within their EEZs. Prior to public release, all data will be considered preliminary. Such data will be available to participating scientists, who should consult the data originator about its status. Data should be shared with other cruise/process study participants as soon as they become available during or after a cruise or process study, to enable data synthesis to proceed rapidly, with the understanding that the data are the proprietary material of the originating scientist and may not be used without their permission. However, for non-participating scientists the data can be obtained only with the permission of the responsible participating scientist.

The recipient DAC will not publicly redistribute such data, or a derivative containing most of the information during the publication rights period.

The receiving investigator should not publish any paper based predominantly on the received data during the publication rights period, should co-author results with the originating investigator, and should not redistribute the data.

Adherence to this data policy is expected of all scientists participating in national and international GEOTRACES activities.

5.5 Data centres

GEOTRACES should work with one of the Earth sciences data centres to develop a GEOTRACES data portal, providing integrated access to GEOTRACES data through the Internet. Access through the portal should allow for flexible, ad hoc queries and data downloads to common formats; and both public and private access, which should be independent of the location of the data. A security layer is needed to allow access to non-public data for PIs with permission, including capability for groups of PIs to access subsets of the data, based on approved user lists and passwords. Permissions will have to be managed by the DLO, according to policies developed by the DMC and approved by the SSC. The portal, and its underlying cyber-infrastructure, should also adhere to GEOTRACES data policies regarding data availability, proprietorship and release.

The DLO should maintain a catalogue of all data available, including metadata types. The portal should make the catalogue, with contact information, easily available to facilitate data access. The catalogue will encourage collaboration within GEOTRACES, in addition to timely linkages with other research programs (e.g., IMBER, SOLAS, GEOHAB, LOICZ).

The portal should be supported by a relational database housing the GEOTRACES water column data and any of the GEOTRACES data streams that are not housed at other already-existing data storage and access facilities. Data streams that are housed at other data centres (e.g. CCHDO, see below), along with their relevant metadata, should also be accessible through the portal. The database should be structured so that metadata can be accessed through the portal, either separately from or along with the data. A metadata report should be available that offers users a list of the metadata available for each of the data categories. Data downloads should always include at least the version, units, quality

and citation metadata for each dataset. Preliminary data, not yet publicly released, should be accompanied with a message clearly stating as much.

The GEOTRACES DAC should provide data support services for nations or smaller GEOTRACES programs that lack their own data support infrastructure. Support would be required from the DAC for formatting and transmission of measurements and metadata.

The GEOTRACES data portal should accept and encourage feedback from users regarding technical and data quality issues. The DAC should be responsible for technical access issues. Issues related to data quality should be forwarded by the DAC to the relevant PI, with a copy to the GEOTRACES DLO. The IPO should track that data quality issues are addressed by the PIs.

Water column data that are compatible with the CLIVAR hydrographic data types should be submitted to the CCHDO at Scripps. The water column data are the central data set of GEOTRACES, so GEOTRACES will benefit greatly from the expertise of CCHDO. In addition, many of the GEOTRACES hydrographic data sets are likely to be part of a continuum of data sets from other programs, e.g. CLIVAR, WOCE etc. Thus it is important that these aspects of the GEOTRACES data sets be quality controlled, stored and archived in a seamless manner with the other global data sets. However, the CCHDO does not have the capability to handle the wide range of data types that GEOTRACES will collect, so cannot be the overarching GEOTRACES DAC. The CCHDO also does not hold its data in relational databases, which GEOTRACES would like to make use of to facilitate data integration across cruises and between distinct data types. Thus, an overarching data centre must be developed. The water column data submitted to the CCHDO will be accessible (for example by mirroring) in the GEOTRACES database.

5.6 Special requirements of process studies

A simple, convenient data policy and formatting system is the best way to ensure international agreement and will facilitate and promote regional cooperation and collaboration. Nations without oceanographic data centres and/or lacking funds to build and maintain data management facilities will then be encouraged to submit their data sets to the international data centre.

- Regardless of data storage details, the process study results should be distributed through the same portal as the core activity data.
- End user data retrieval for process study results should have the same interface (look and feel) as core data.
- Process study results should use the same data stream processing facilities as used for core activities whenever possible.
- Established data processing facilities should be used rather than building new capabilities whenever possible.
- Process study metadata (who, what, when, where, how) should be submitted to the GEOTRACES data centre as soon as soon as possible (within one month after completion of the field work or each phase of the field work) using either a detailed preliminary project report or standardized metadata forms specifically designed by the data manager for the specific process study.

- A final detailed data report is required for each process study. These reports will be similar to a cruise report. In addition to describing the process study, these reports will provide one mechanism to give credit to data generators. The final reports will be publicly available, distributed through the GEOTRACES portal and, if possible, be assigned a reference DOI.

5.7 Timing and cost

The first GEOTRACES section data will be available in 2007, as part of the GEOTRACES IPY cruises, so it is important to have the data management system in place well before then. Some cruise-related metadata are already available. At its first meeting in 2006, therefore, the GEOTRACES SSC will need to establish the Data Management Committee, and develop proposals to fund GEOTRACES data activities. As soon as the location of the IPO is determined, the Data Liaison Officer should be appointed, hopefully also in 2006. GEOTRACES should begin to develop relationships with existing and potential DACs in 2006.

Previous projects have shown that the proportion of the total project science budget (including platform costs) required for end-to-end project data management, project data services, and assuring the long-term stewardship of the project data is approximately 10%.

6. Recommendations

1. GEOTRACES should adopt, as much as possible, successful approaches used by other projects.
2. GEOTRACES should establish a Data Management Committee (DMC) comprised of data originators (i.e., observational scientists), data managers (national and international) and data users (including modellers).
3. A Data Liaison Officer (DLO) should be appointed to work at the GEOTRACES International Project Office (IPO). The primary responsibility of the DLO will be to define and maintain metadata for GEOTRACES. The DLO should be appointed as soon as the IPO is established.
4. The GEOTRACES SSC should establish the DMC and the IPO and appoint the DLO in 2006, and begin to develop relationships with existing and potential DACs in 2006. All of these require that funding be sought.
5. GEOTRACES should make use of the CLIVAR and Carbon Hydrographic Data Office (CCHDO), which is already funded to take in profile data internationally, but should arrange another DAC to develop a relational database to serve all GEOTRACES data.
6. GEOTRACES should work with one of the Earth sciences data centres to develop a GEOTRACES data portal, providing integrated access to GEOTRACES data through the Internet.
7. It is strongly recommended that data management professionals be involved in all GEOTRACES data activities from the start. This is known as co-operative or “end-to-end” data management. Practically, this should be achieved by the assignment of a Data Specialist to each cruise or process study (in addition to the

DLO at the IPO), whose primary role is to assist the lead scientist in metadata collection and data management. Allocation of funding (from a PI's grant or a Data Centre) will pay off amply in the completeness and quality of the GEOTRACES final data sets.

8. Endorsement of scientific activity by GEOTRACES requires metadata to be submitted on the shortest possible time scales. Failure to do so should be considered reason to remove GEOTRACES endorsement.
9. GEOTRACES should adopt the Directory Interchange Format or a similar discovery metadata standard, for example, ISO19115 when it is completed.
10. Metadata should be delivered to the IPO as soon as created, from the planning stage onwards. Data (not finalized) should be submitted to a DAC within 1 month (of collection, or end of cruise), with possible approved extension as tabulated. Cruise or project reports should be submitted to the IPO within 6 months of the end of the cruise or process study. Final data, following international guidelines, should be submitted within 2 years of cruise or process study completion, with exceptions possible from the GEOTRACES SSC.
11. Data should be shared with other participants in a particular cruise or process study from an early stage (as soon as it available at the DAC, with knowledge and permission of the relevant PI)
12. Public release will normally be two years from the end of the cruise or field activity, but should be extended when analytical procedures have inherent built-in delays.
13. Individual PIs are responsible for quality control of their data. Groups of PIs expert in a particular parameter will be encouraged to apply further quality controls.
14. The GEOTRACES DMC, when approved, should discuss specific data management requirements for GEOTRACES process studies, when it is better known what these studies will be.
15. The GEOTRACES SSC should consider providing access to project-related publications through a publication database, such as that used by GLOBEC.
16. The World Data Center for Oceanography, Silver Spring, should be considered as the most appropriate WDC for GEOTRACES data. It is recommended that the DMC contact this WDC to tell them about GEOTRACES once the project is under way, and discuss long-term archiving.

Appendix 1 – list of attendees

Dr Juan Brown
British Oceanographic Data Centre
Joseph Proudman Building
6 Brownlow Street, Liverpool L3 5DA, UK
email: jbrown@bodc.ac.uk
Phone: +44 (0)151 7954880
Fax: +44 (0) 151 7954912

Dr Robert M. Key
Atmospheric and Oceanic Sciences Program
Sayre Hall, Forrestal Campus
Princeton University, Princeton, NJ 08544, USA
email: key@princeton.edu
Phone. +1 609-258-3595
Fax: +1 609-258-2850

Dr Roy Lowry
British Oceanographic Data Centre
Joseph Proudman Building
6 Brownlow Street, Liverpool L3 5DA, UK
email: rkl@bodc.ac.uk
Phone: +44 (0)151 795 4895
Fax: +44 (0) 151 795 4912

Prof. Chris Measures
Department of Oceanography, University of Hawaii at Manoa
1000 Pope Road,
Honolulu, HI 96822, USA
email: chrism@soest.hawaii.edu
Phone: +1 808 956 8693
Fax: +1 808 956 7112

Dr. Bob Newton
Geochemistry 63, Lamont Doherty Earth Observatory
PO Box 1000, Palisades
NY 10964-8000, USA
email: bnewton@ldeo.columbia.edu
Phone: +1 845 365 8686
Fax:

Prof. Raymond Pollard
National Oceanography Centre, Southampton
European Way, Southampton SO14 3ZH, UK
email: rtp@noc.soton.ac.uk
Phone: +44 (0) 23 80596433
Fax: +44 (0) 23 80596204

Prof. Reiner Schlitzer

Columbusstrasse
D-27568 Bremerhaven (Building D-1160)
Germany

email: rschlitzer@awi-bremerhaven.de

Phone: +49 (471) 4831 1559

Fax: +49 (471) 4831 1149

Dr James H. Swift

UCSD Scripps Institution of Oceanography, Mail Code 0214

9500 Gilman Drive

La Jolla, CA 92093-0214, USA

email: jswift@ucsd.edu

Phone: +1 858 534 3387

Fax: +1 858 534 7383

Dr Marie-Paule Torre

Observatoire Oceanologique, Base de données PROOF

Cas Nicolas, Quai de la Darse, BP 8

06238 Villefranche sur Mer, France

email: torre@obs-vlfr.fr

Phone: +33 4 93763877

Fax: +33 4 93763873

Dr Ed Urban

SCOR Secretariat, Dept of Earth and Planetary Sciences

The Johns Hopkins University

Baltimore, MD 21218, USA

email: Ed.Urban@jhu.edu

Phone: +1 410 516 4239

Fax: +1 410 516 4019

Dr Jing Zhang

Faculty of Science, University of Toyama

Gofuku 3190

Toyama, 9308555, Japan

email: jzhang@sci.toyama-u.ac.jp

Phone: +81 764456665

Fax: +81 764456549

Appendix 2 – detailed list of data types

Geochemical data types--Ocean Sections

Abbreviations and notes

At this stage no detailed implementation plan exists for GEOTRACES. The table below has been compiled based on information in the GEOTRACES Science Plan and is for data planning purposes only. It is not intended to be proscriptive and is subject to change by the GEOTRACES SSC.

Each TEI parameter is expected to generate at least one value - total concentration. In many cases, for each TEI will be generated values for total and dissolved concentrations,

and potentially more than one version of the dissolved value as a result of the use of different filter sizes (for example colloidal materials may be collected).

In the case of some TEI different chemical speciation forms will be reported. For example, selenium (Se) will be reported in three chemical forms: +VI, +IV and organic Se. Elements where this is likely have a Y in the multiple forms column.

Only relevant boxes are marked, no mark means not applicable, or samples are not expected to be taken.

Sampling

1 db = continuous sensor signal averaged to 1 or 2 decibar pressure intervals.

1 minute = continuous sensor signal averaged to 1 minute or similar time interval

A = all, i.e. it is expected that a sample will be collected from each bottle fired, to check sample integrity and provide comprehensive background information.

H = Samples expected to come from hydrography rosette.

K = Key parameter expected to be sampled on each GEOTRACES section, but not necessarily at all stations

S = some sampling, but less than total sampling.

T = Samples expected to come from trace metal rosette.

WA = where appropriate. On some cruises, where needed and shipboard space/water volumes permit.

Y = sampling expected

? = sampled form/rate undecided

Parameter	Sampling	Total	Dissolved	Multiple forms	Isotopes	Comments
Routine parameters at all stations, all cruises						
Hull and deck mounted underway sensors						
Atmospheric	1 minute					
Surface ocean	1 minute					
Multibeam	1 minute					
pCO ₂	?	Y				
Rosette mounted sensors						
Temperature	1 db					
Salinity	1 db					
Dissolved O ₂	1 db					

Chlorophyll	1 db					
Beam transm.	1 db					
Water samples						
Sensor validation, water sample integrity, hydrographic setting						
Salinity	A	Y				
Oxygen	AH	Y				
Nitrate	A	Y				
Nitrite	A	Y				
NH ₃	A	Y				
Silicate	A	Y			YS	
Phosphate	A	Y				
Key parameters, sampled on each section but not each station or bottle						
Trace metals						
Fe	KT	Y	Y	Y	YS	
Al	KT	Y	Y	Y		
Zn	KT	Y	Y	Y	YS	
Mn	KT	Y	?	Y		
Cd	KT	Y	?		YS	
Cu	KT	Y	?			
Stable isotopes						
¹⁵ N	KH	Y				
¹³ C	KH	Y				
Radioactive isotopes						
²³⁰ Th	KH	Y			Y	
²³¹ Pa	KH	Y				

Radiogenic isotopes						
Pb	KT	Y	?		Y	
Nd	KH	Y	?		Y	
Other parameters						
Stored samples	KT	Y	Y		Y	
Particulates	KT	Y				
HPLC pigs	KH	Y				0-200m
Aerosol TSM	K	Y				1 d integration
Parameters measured on some, but not all, sections						
Carbon parameters						
TCO ₂	AH	Y				
pH	AH	Y				
Alkalinity	AH	Y				
Trace metals						
Fe species	ST	Y	Y	Y		
Se	ST	Y	Y	Y		
Ni	ST	Y	?			
Ge	ST	Y	?			
Ti	ST	Y	?			
Co	ST	Y	?			
Ga	ST	Y	?			
REE	ST	Y	?	Y		
Ba	ST	Y	?		YS	
Mo	ST	Y	?			
V	ST	Y	?			
Hg	ST	Y	?	Y		
Sn	ST	Y	?			

Ag	ST	Y	?			
Stable and radioactive isotopes						
Ac	SH	Y				
Ra	SH	Y			Y	
Be	ST	Y			Y	
Pb	ST	Y			Y	
Hf	SH	Y			Y	
Nd	SH	Y			Y	
Rn	SH	Y				
Physical tracers						
CFCs	SH	Y				WA
³ H- ³ He	SH	Y				WA
Inert gases	SH	Y				WA

Process studies

It is impossible to specify parameters for process studies in detail, as too little is known of what is likely to be proposed. In addition to the parameters listed for ocean sections, there will certainly be, for example, bottom landers; sediment traps; sediment cores; time series from moored and drifting buoys; and net tows, to name but a few.

Appendix 3 – individual presentations

A3.1 GEOTRACES scientific overview – Chris Measures

The GEOTRACES program will consist of ocean sections and process studies. Sections will be designed to cross property gradients that are of interest to GEOTRACES (see above), and process studies will be designed to investigate rates of processes. The initial part of the program will consist of ocean sections with process studies occupying the middle and later part of the program. The exact cruise tracks for ocean sections will be determined at one or more meetings of interested scientists (see GEOTRACES Science Plan for map of important areas for ocean sections and process studies).

Ocean Sections

Ocean sections are expected to consist of vertical water sampling at a station spacing of several degrees. In addition there will be less-frequent stations at which multiple casts will be conducted. In addition to the vertical profile bottle sampling at discrete stations, the following data-generating activities will occur:

- Continuous aerosol sampling (this sampling may be conducted by SOLAS scientists on GEOTRACES cruises)
- Underway pCO₂, nutrients
- In situ pumping systems
- Gathering of routine meteorological and hydrographic parameters, including surface underway parameters (IMET), temperature, salinity, nutrients.
- Shipboard ADCP, Multibeam

Additionally, in remote regions, a limited amount of coring (gravity, box, piston) will be undertaken. Net tows also will be undertaken in a limited number of places.

Water column sampling data streams

Water sampling will be conducted using a 24- or 36-place “hydrography” rosette and a 24-place Trace Metal clean rosette. It is anticipated that all the routine hydrography parameters will be derived from the hydrography rosette. Remaining water from the hydrography rosette will be used for determination of trace elements and isotopes (TEIs) that are not sensitive to contamination from standard hydrography bottles or rosettes.

The Trace Metal rosette will be used for contamination-sensitive TEIs. Additional samples for salinity and nutrients (in some areas, nutrients are much better “trip indicators” than salinity) will be drawn from this rosette to verify correct firing of the Go-Flo bottles and to provide a detailed hydrographic context for these samples.

At large-volume stations, multiple tripping of sample bottles and multiple casts of the rosettes will be used to obtain the larger volumes required for determination of some TEIs. In addition to multi-parameter data (see below) generated from water samples, each rosette system will be instrumented, thus producing continuous data that will be averaged into 1-2 db intervals. Sensors on each rosette will usually measure conductivity, temperature, pressure, dissolved oxygen, fluorescence, and beam attenuation.

Process studies

Process studies will be conducted during and after the ocean-sections component of the programme is completed. Process studies will be conducted in regions and locations that are chosen to study specific processes that control ocean TEI levels and distributions. The study regions may already be obvious or may be identified through the ocean sections. The following types of measurements are likely to be conducted at process study sites:

- Water sampling at high resolution spatially (short ocean sections of bottle data and data from towed sensors) and temporally (time series from moored and

drifting buoys, and periodic re-occupation of long-term observation sites, such as BATS, HOT, KNOT, etc.)

- Bottom landers for measuring benthic fluxes
- Sediment traps to estimate particulate fluxes
- Sediment coring to conduct ground truthing of TEIs thought to be useful for paleo-proxies of oceanographic conditions. The coring will focus on obtaining good sediment-top samples.
- Net tows to collect organisms used as paleo-recorders, such as foraminifera, to calibrate these recorders with water-column and sediment TEI concentrations
- Remote sensing

A3.2 Hydrographic and nutrient data – Jim Swift

The water column measurements carried out during GEOTRACES transect cruises, although focusing on TEIs, must include determination of those routine and globally well-known physical and chemical parameters (temperature, salinity, pressure, oxygen, nutrients) appropriate to place the TEIs into the broad oceanographic context. These “routine” hydrographic data, as they are sometimes known (although some require great effort to acquire), must be of reference quality in order to both facilitate cruise-to-cruise data assembly and comparison, and also to contribute to a long service life for GEOTRACES measurements. Fortunately, one of the legacies of the World Ocean Circulation Experiment (WOCE) Hydrographic Program (WHP) is infrastructure and experience in the acquisition and processing of reference-quality routine hydrographic data well matched to GEOTRACES requirements.

The basic suite of routine hydrographic measurements and associated data strongly recommended for all GEOTRACES transect cruises includes station and cast summary information (“headers”), CTD profiles, CTD oxygen sensor profiles, CTD fluorometer and transmissometer profiles, and rosette water sampling, including salinity, oxygen, and nutrient determinations.

Station and cast summary (“header”) information

The basic unit of water column data is the “cast”, a single deployment of a rosette with empty bottles, firing of the bottles at desired depths, and recovery of the rosette and full bottles. Each GEOTRACES “station” will include one or more casts. For each cast, cruise scientists must report, at a minimum, the cruise name or expedition code, station number, cast number, cast position, and cast time. Additional cast header information (desirable but not required) includes ship name, GEOTRACES transect identifier, cast type, and depth to bottom (very highly recommended but not required).

Rosette sampling (bottles)

Each GEOTRACES water sample will be collected in and drawn from an oceanographic sampling bottle, usually from rosette casts, but occasionally from wire casts. Water sampling bottles can contribute to sample contamination. Thus, each bottle should be uniquely numbered and tracked, at least within each cruise. (It is optional to carry over bottle tracking between cruises.) Modern rosette bottle closure controllers, such as those from SeaBird, are usually well-behaved and reliable. Nonetheless, the water sample data

should be examined in relation to the CTDO (CTD + oxygen) data to see that bottle closures took place as intended by the CTD operator. Water sample bottles are subject to leaks, sometimes frequently and/or severely. Thus, it is very important that the first individual to draw a sample from a bottle carry out a standard leak test, and that any discrepancies be noted at once on the sample log sheet. Each bottle closure in a data file should have a quality code associated with it. A leak later found to have adversely affected the parameter data should result in the bottle quality code being changed from the code for “good” (the default) to that for “bad.” If it remains unclear if the leak adversely affected the data, the bottle quality code should be changed to the value for “questionable.”

Note that due to the need to eliminate the effects of rosette wake, a rosette must be stopped for at least 20 seconds before bottle closure. An alternative technique, which avoids contamination by the rosette wake, used principally during TM rosette work, is to close the rosette bottles “on the fly”, that is, without stopping but at a reduced (5-10 m/sec) haul speed. But, in either case, the inherent flushing length of the bottle must be considered versus the vertical resolution required for all the parameters to be measured from the bottle. In most open-ocean conditions with moderate-to-light ship roll, waiting two ship-roll cycles before closing the rosette bottle will provide adequate flushing. In extremely calm situations, it is sometimes desirable to mimic one or two ship rolls by “yo-yoing” the winch a meter or two up and down at an appropriate rate.

One cannot determine accurately later what samples were drawn from each bottle. Thus, every sample must be documented on a sample log sheet when it is drawn, for example, by recording its serial or container number. This recording is typically carried out by a person designated as the “sample cop,” who also guides samplers to available bottles in the appropriate sampling order. A very important function of this person is to record bottle information and other sampling comments made during sampling, so that the data from suspect bottles or samples can be checked at sea and ashore during data processing and quality control.

CTD

It is very strongly preferred that each GEOTRACES water column profile include a CTD cast, preferably a full-depth CTD profile. The CTD data are typically reported as a processed pressure series at an appropriate pressure interval, for example, one set of values for each two decibars from the surface to the bottom.

The CTD sensors should be maintained and calibrated in accordance with the manufacturer's recommendations, and this information should be part of the metadata. The manufacturer's pressure and temperature sensor calibration are adequate for GEOTRACES. The manufacturer's CTD data acquisition software is adequate for GEOTRACES, though custom acquisition software offers some advantages.

The CTD data processor must fully document the processing. The manufacturer's data processing software, in experienced hands, is adequate for GEOTRACES but can be improved. The experience of the data processor is more important than the software package. It is very useful to have a CTD data specialist at sea. CTD data validation and quality control are the responsibility of the data processor, working with a scientist familiar with CTD data and the oceanographic domains that were sampled.

Each GEOTRACES water sample must have associated with it pressure, temperature, and salinity values. For data from most rosette casts, these data will come from a CTD attached to the rosette, typically an average of processed CTD values over an appropriate interval very close to the time the rosette bottle was closed by the CTD operator. In rare cases some GEOTRACES water samples may be collected from bottle or rosette casts with no CTD device available. In such cases, the scientists responsible must develop a scheme to obtain pressure, temperature, and salinity values for each bottle closure of appropriate utility and quality. For example, if the salinity gradient is strong enough, salinity-check samples from each of the bottles may be adequate to relate the bottle sample to a CTD profile as nearly coincident as feasible. The practice of estimating bottle depths, and interpolating CTD data, can result in difficulties, although interpolation may be necessary if no information of greater reliability can be obtained. It is very important that the cruise and cast documentation very clearly identify the methodology used in these non-standard situations.

CTD oxygen sensor

Use of a CTD dissolved oxygen sensor can provide useful, unique information on vertical structure of dissolved oxygen. Experience has shown that the SBE-43 sensor represents a significant improvement over previous CTD oxygen sensors. The oxygen sensor cannot be laboratory calibrated except for casual work. Hence, correction to water samples is required for any serious use. This presents difficulties because oxygen sensor hysteresis (and also flow-rate changes when the winch is stopped for bottle closures on the up cast, plus sensor orientation and wake issues) complicate obtaining a CTD dissolved oxygen value to associate with a given rosette bottle sample, at the quality expected for typical scientific work. Therefore, the usual practice is to obtain a bottle oxygen sample from every level, and match down-cast sensor oxygens to up cast bottle oxygens matched by corrected density. Note that bottle oxygen measurements are not a trivial addition to a cruise. Also, CTD dissolved oxygen sensors have significant co-response (i.e. to temperature, flow rate) plus sensor shifts, drifts, and aging, and sensitivity to cold air, so that significant expertise is required in the processing, evaluation, and documentation of CTD dissolved oxygen sensor data.

Other sensors on the CTD

A CTD package may include additional sensors and devices other than those mentioned above. Typical examples include an altimeter, a transmissometer, and one or more fluorometers. Some of these instruments may record internally, but typically these add to the CTD data stream. It is crucial that the CTD team be notified many months in advance of a cruise so that appropriate arrangements and modifications can be made to accommodate any extra data channels required. For example, the CTD digitizer might handle only limited extra data channels, or the total power available may limit some sensors. Further, the details of collection, correction, and data processing are often overlooked by seagoing teams. Therefore, to the degree that data from extra sensors are crucial to GEOTRACES science, appropriate care must be taken in preparation, collection, processing, and reporting of such data.

Salinity (bottle)

Collection, analysis, processing, and reporting of discrete (bottle) salinity samples to WHP standards is adequate for GEOTRACES. It is necessary to follow established protocols, use suitable sample containers, and store samples appropriately. Typical problems are contamination from freshwater droplets or condensate, salt crystals, or evaporation; salinometer stability; inadequate experience of the salinometer operator; and timely recognition of suspect data. Proper application of IAPSO Standard Seawater is essential. For GEOTRACES, a few high-quality samples for CTD correction are better than a large number of samples of mediocre quality. Note that with the present SeaBird conductivity sensors and appropriate care in CTD acquisition and data processing, the primary salinity data can come from the CTD data at bottle closure time, and thus the bottle salinities can be principally only those needed for absolute calibration of the CTD data, and evaluation of sample integrity, which is a required check for Go-Flo bottle sealing. It is required to report the IAPSO Standard Seawater batch number and of all analyses that fall outside norms. Comparison of discrete salinity values to CTD values is useful for assigning quality codes.

Oxygen (bottle)

Obtaining discrete (bottle) oxygen values at appropriate quality (WHP standards) is time consuming, requiring care and expertise at every stage: pre-cruise preparation, reagent preparation and monitoring, standardization, sampling, analysis, data processing and quality control. The WHP protocols, in experienced hands, are adequate to GEOTRACES needs. There are possible benefits to GEOTRACES to have program-wide oxygen standards for all cruises provided from one laboratory. Comparison of bottle oxygen values with CTD oxygen is useful, but not definitive for data evaluation, and it is preferred to evaluate oxygen data with those from other routine hydrographic measurements.

Nutrients

The GEOTRACES plan requires measurement of silicate, phosphate, nitrate, nitrite, and ammonium (abbreviated here as SiO_3 , PO_4 , NO_3 , NO_2 , and NH_4) on ocean sections. Although collection of nutrient samples is uncomplicated, analysis of this suite of parameters to high international standard (e.g., WHP standards) can pose difficulties for all but a relatively small number of laboratories. Various analytic equipment and protocols are in use, some with a high degree of automation. Individually, these measurements may appear to produce internally consistent results. But real-world experience during WOCE yielded 1-3% concentration differences between well-respected laboratories. These differences point to unresolved or not widely recognized technical and standardization issues. A potentially more significant problem lies in the absence of reliable and effective international reference standards for nutrient data. It may be useful for GEOTRACES to hold an international nutrient workshop involving the technical teams that will run GEOTRACES nutrients. Data validation and quality control can be conducted to some extent by the cruise data specialist, but there is also a need to merge nutrient data with other routine hydrographic data and co-evaluate the data.

Final comments

- It is useful to have on-board data processing specialists for both CTD and bottle data.
- Scientific use of data provides some of the best quality-control information.
- It is very important for the data center(s) to know soon: Do GEOTRACES investigators require access to a true central database of the parameter data (as opposed to a *much* simpler database of metadata)?
- Parameter units are an extremely important issue (nutrients are an example of a serious case) that should be specified as part of the data system description.
- WOCE Hydrographic Program standards are appropriate to GEOTRACES, though it should be noted that they are not as well defined as they might be.
- Sample log sheets are crucial. Each sample drawn from a sample bottle must be logged.
- There are many other considerations that potentially affect the relationship of the measured parameters. See, for example, "Data Evaluation and Quality Control for Routine CTD/Hydro Data" <http://www.gso.uri.edu/unols/inmartech98>.

A3.3 Data management overview – Roy Lowry

For definitions of data management technical terms, see Appendix 4.

Data Management Issues

Data management issues include project dataset scope, description, instantiation (for definition see Appendix 4), and quality assurance; interoperability of databases and data standards; project data policy; intra-project data exchange; data management infrastructure; and long-term stewardship of the data.

Project Dataset Scope

The project dataset is an aggregation of datasets from different project activities. Building a catalogue of these activity datasets is the most basic task of project data management. However, it can be far from straightforward. There can be controversy about what comprises a project cruise. For example, Belgica OMEX cruises were considered as totally unrelated to JGOFS by some people and totally contributory to JGOFS by others. Scientific Steering Committees have trouble saying no to willing contributors of data and have a hard time constraining the project scope.

There can be over-enthusiasm at the national level for activities to include in the project dataset, even if peripheral to the project objectives (one nation submitted several hundred cruises to the JGOFS database, most of which were of peripheral relevance). Some project activities are shared cruises with other projects, creating data ownership and access issues. For example, who should have access to CTD datasets from WOCE cruises carrying a JGOFS team doing carbonate system measurements? It is imperative that such issues are addressed by the responsible SSCs and IPOs before cruises take place.

Project Dataset Descriptions

Dataset catalogues need to be more than just a list of dataset names. Each dataset needs to be described by a discovery metadata record, which provides data that makes it possible to find the data based on relevant descriptors. According to the 2003 SCOR/IGBP data management meeting (<http://www.jhu.edu/scor/DataMgmt.htm>), discovery metadata should be compiled and managed by the IPO.

Project Dataset Instantiation

A collection of data interchange formats (DIFs) needs to be converted into a collection of data files that make up the datasets described. This process requires careful planning and management during the life of the project.

Project Dataset Quality Assurance

Every project must determine who is responsible for project data quality.

PIs can be given responsibility for specialist parameters. But who ensures that all CTD parameters are fully worked up and calibrated? And who is responsible for metadata quality assurance? These issues must be worked out by the GEOTRACES Data Management Committee.

Interoperability and standards

Data standards are necessary to ensure interoperability of databases. WOCE had strong syntactic and some degree of semantic data standards, but JGOFS did not. As a result, there was an integrated WOCE dataset available at the end of the project, but an integrated JGOFS dataset needed 5 years of post-project work to produce.

The value of data and metadata interoperability cannot be overstated. Interoperability can be achieved most easily through universal adoption of standards. Syntactic interoperability comes easily through adoption of mature technologies (e.g., netCDF). Semantic interoperability is much harder to achieve. Content standards and controlling vocabularies for soft-typed elements are essential to achieve semantic interoperability.

Data policy

A project data policy specifies who can have access to what data and when. The policy specifies who gets what reward for what actions. Simple policies are the simplest to implement. Liberal policies also are simple to implement, but may not offer the desired features. The project data policy needs to be specified and agreed at the outset. As part of the policy, the issue of data release to the public domain needs to be addressed

The data policy also needs to specify how project participants obtain project data held by other participants. Today, anything other than Web-based data access is inconceivable. The intra-project data exchange expected within GEOTRACES has infrastructure implications.

Major functions of a data policy are expectation management and policing the consensus.

Infrastructure

The design of the project data infrastructure has many issues that should be considered:

- Will the system have visualisation/usage tools? The answer to this question may influence other decisions, such as standards. One of the valuable features of the CLIVAR data management system on cruises is the availability of data-plotting software that combines cruise data in near-real time.
- To what extent will the project data management system be a distributed system versus a centralised system? The desirable granularity is the crucial issue. Too fine granularity (e.g., each PI's computer as a node), results in poor reliability. A coarser system requires ingestion of data from distributed data origination sites and therefore resources to accomplish this integration. A single centralised system may be unrealistic, due to the resources required.
- Will a DAC concept be used for designated data types? This approach worked well for WOCE. DACs for specific data types can be shared with other projects, which is an important consideration as other new marine research projects are developing their data management systems.

Long-term stewardship

Projects must consider long-term stewardship of project data after the project is completed and the IPO is closed. Will the data be published on physical media? Will the project maintain a Web presence on an established server after the end of the project? The JGOFS Web site has continued on the University of Bergen server for two years after the JGOFS IPO closed (so far), but it is unclear how long this university's server will host the Web site. JGOFS data DVDs (Vol. 2) will be available through WDC-MARE. For longer-term data storage, project data should be integrated into an established distributed system, specifically, one or more World Data Centres (WDCs), depending on the data type. It is very helpful for projects to involve the appropriate WDCs early in the project, rather than merely expecting to dump the data into the WDCs at the end of the project with no warning.

A3.4 Data Centre interest in GEOTRACES

BODC - Juan Brown

BODC (www.bodc.ac.uk) is the United Kingdom's national marine data centre and holds national archives of sea level, current meter, hydrographic and a considerable quantity and range of multidisciplinary data (chemical, biological, geophysical and physical), with databases containing almost 10,000 parameters. BODC is primarily funded by the UK's Natural Environment Research Council (NERC), with a responsibility to ensure the long-term security and delivery of UK marine data (quality assured).

A significant proportion of the data held by BODC are (and have always been) derived from outside NERC, for example, from government departments, industry and BODC's management of large-scale European (e.g., MAST OMEX and PROVESS) projects. The guiding principle for BODC data management is that oceanographic data are temporally and spatially sparse, so only through an amalgamation of national resources can the data's true worth be realised. Consequently, NERC-funded scientists gain ready access to a much wider pool of data and information.

BODC specialises in multi-disciplinary project work, such as proposed by GEOTRACES (see, for example, <http://www.bodc.ac.uk/projects/uk/amt>). BODC data are held in a

relational database providing a powerful search capability across multiple parameters, projects and cruises. BODC employs PhD-trained marine scientists who understand the data, and why they are being collected and used. BODC staff members frequently accompany cruises to ensure delivery of quality data. When required, BODC staff work up data and always apply quality control before banking and making the data available for the long term.

BODC represents the United Kingdom in terms of marine data management at UNESCO's Intergovernmental Oceanographic Commission (IOC), the International Council for the Exploration of the Sea (ICES), and European Union-sponsored projects, and acts as the CLIVAR centre for data from moored instruments and sea level sensors.

BODC Interest in GEOTRACES

BODC will be responsible for UK-funded GEOTRACES projects. Whilst BODC would welcome the opportunity to be more widely involved in GEOTRACES, particularly as BODC regards dealing with such programmes as one of its strengths (see above), the BODC funding model requires that additional funding be supplied for management of international project data. BODC staff would be happy to discuss the possibility of a wider role in GEOTRACES.

CCHDO - Jim Swift

The CLIVAR and Carbon Hydrographic Data Office (CCHDO) at the UCSD Scripps Institution of Oceanography serves the oceanographic community by locating, collecting, inspecting, modifying to improve adherence to standards, organizing, and making available the CTD, hydrographic, tracer, and ocean carbon data and associated documentation relevant to studies of the large-scale circulation and water masses of the world ocean. The CCHDO is a continuation of the WOCE Hydrographic Program (WHP) Office, and as such ensures that any data submitted to it conform to WOCE and CLIVAR standards. The CCHDO's primary presence is via the Web site <http://cchdo.ucsd.edu>, which contains the CCHDO's complete public holdings, documentation, data submission guidelines, and so forth.

The CCHDO is an appropriate home for water column data generated during GEOTRACES, including CTDO profiles and all water sample parameters. For example, the CCHDO can readily deal with the routine hydrographic data that are intended to be made public shortly after each GEOTRACES transect cruise, and can update these, and/or manage GEOTRACES TEI data as they become available for distribution. CCHDO staff can work with the GEOTRACES IPO to locate cruises whose CTD and water sample data should be at the CCHDO, arrange transmission of those data to the CCHDO, report back to GEOTRACES on progress, and add GEOTRACES data to the office Web site. If GEOTRACES conducts only a small number of cruises each year, and if the CCHDO has the opportunity to work with GEOTRACES investigators before, during, and after cruises to facilitate interactions, the additional funds required will be principally those to support travel of CCHDO staff to GEOTRACES meetings, as appropriate. The CCHDO is presently funded through 2008, and as long as the research community continues to utilize the CCHDO, renewals will be sought.

Lamont-Doherty / CIESIN Capabilities – Bob Newton

The Center for International Earth Science Information Network (CIESIN) and the Lamont-Doherty Earth Observatory (LDEO) would like to propose development, maintenance and hosting of the GEOTRACES DAC at CIESIN in Palisades, New York, USA. CIESIN/LDEO would propose creation of an integrated GEOTRACES portal for all data access, using a relational database for data storage, and using current-generation data management utilities. CIESIN would be willing to build the GEOTRACES portal and databases to the program's requirements, as expressed by the GEOTRACES Data Management Committee and Scientific Steering Committee. Development of the database is estimated to require roughly 3 person-years, which could be accomplished in 6 to 12 months, depending on the complexity of the effort. Once created, it is estimated that the ongoing maintenance and operation of the GEOTRACES DAC would require approximately 50% of a position. The rationale for this approach (a dedicated portal and relational database) is twofold. First, it will provide GEOTRACES PIs and data consumers with a research tool that implements the central goal of the program, an integrated assessment of diverse trace elements and isotopes. Second, it will save large quantities of resources, as the data integration across variables and with all program metadata will be accomplished in a standard, transparent and commonly accessible way.

A bulleted summary of CIESIN's relevant expertise and capabilities appears below:

CIESIN Mission:

- To create information resources and knowledge networks for science and decision making from local to global scales
- To promote understanding of sustainable development and the interactions between humans and the environment.

Core CIESIN Programs:

- One of 8 DAACs in the NASA Earth Observing System Data and Information System
- Interactive services including: Ramsar DG, ENTRI, US-Mexico DDViewer, GISS Crop-climate, ESI Viewer.
- Data Products including GPW V2, ESI, Last of the Wild/Human Footprint, PLACE, SRES
- Geochemical databases:
 - PetDB <http://www.petdb.org>
 - SedDB <http://www.seddb.org>
 - SESAR <http://www.geosamples.org>
 - EarthChem <http://www.earthchem.org>

Geochemical Databases Technical Infrastructure Services:

- System development
- System operations
- Data management
 - Scientific data stewardship integrated with researchers
 - Expertise in geospatial data
- Data archiving and long-term archiving based on OAIS approach
- Standards-compliant metadata development, management and catalog support
- User Support Services

- Content management and authoring through open source and commercial tools

Hardware:

- 30 SUN/Solaris Servers
- Storage Area Network (SAN); 4 Tb capacity, expandable to 21 Tb.

Security:

- NASA/NIST compliant IT Security, Risk Management and Contingency plans
- Regular scanning and patching
- NASA ESDIS IT security audit every 6 months
- Firewall protection

System Backup:

- NASA-approved back-up and retention plan
- Legato software with dedicated backup libraries and routine testing
- Daily backups of all development and production servers
- Backup, retention and tape rotation schedule. Annual back-ups are retained indefinitely
- Copies of backup tapes (quarterly, semi-annual, annual) are stored off-site at the NASA GISS facility

Software:

- Oracle 10g Database Server
- OGC-Compliant IONIC mapping of server and client
- ArcIMS Mapping server and client
- WebLogic Application Server

Continuity Services:

- Uninterrupted Power Supply backup for up to 30 minutes
- Generator backup for up to 10 hours (longer with refueling)
- Redundant physical network links
- Standby application, mapping and database servers for failover
- Offsite recovery available as an option.

Personnel:

- 9 full-time information technology professionals
- 12 full-time Science Applications professionals

French PROOF Database – Marie-Paul Torre

The PROOF database, an intermediate structure, is working for a national program: PROOF.

SISMER is the National Data Centre handling PROOF data. A meeting was held on 25 Nov. 2005 in Brest, France about oceanographic databases.

The experience of JGOFS demonstrated the necessity to provide “secure organization” and a unique dictionary, and to prepare database activity before any cruise.

The aim of PROOF data management is

- To use the same syntax (reference for data set/parameter - name, method, sampling - for all campaigns)

- To present all campaigns in the same way
- To deal with heterogeneous data sets
- To involve scientists significantly

Database organisation

Two levels of database organisation have been defined:

- Cruise and parameters described in a metadatabase (DB); meta-information that describes the responsibility for each parameter, access to data sets—from originators—through Web site directory
- More specific: used for discrete data sets (stored in a relational DB in the form of tables): CTD-Rosette - Routines for uploading data sets are not implemented, although the DB committee has requested such routines.

Cruise organization

Before the cruise

- A database correspondent (a scientist) is named and works with the chief scientist. They are in charge of co-ordination and checking metadata and data
- A list of parameters to be measured is compiled
- Cruise metadata are collected

On Board

- DB_correspondent and chief scientist work together

After the cruise

Within two weeks after the cruise, “basic files” are provided:

- List of data measured
- Table of log, events, log stations
- Table of CTD casts references, table of CTD_BTL numbers

Then

- Table of data set expected transmission dates
- Data set not valid/valid/meta-information updated

Cruise participants use an ftp site.

Scientists trust the database because of the organization; datasets are protected and shared by participants within an ftp zone and then through the Web site interface.

Quality Control

- Is conducted by scientists themselves, as a scientific issue
- Is dealt with at the campaign level

A cross validation has been done for cruises. It demands a large knowledge in oceanography, time, with a scientist and a data manager working together.

Dictionary

A parameter is unique because of

- Its name
- Its sampling
- Its method
- Is a stock, flux or ‘-’

Can be a set of data (example: HPLC pigments)

The parameter history document is necessary and updated by scientists.

Participation of PROOF database in the GEOTRACES program

Regarding French GEOTRACES datasets, DB PROOF can store them; the main objective is to conform to GEOTRACES specifications and rules.

In regard to serving as the primary GEOTRACES data centre, this is a political question that needs to be raised with the PROOF_DB committee and responsible scientists.

Appendix 4 – Abbreviations and definitions

ArcIMS	software developed and marketed by ESRI to serve geographic information, such as maps, over the Web (URL). ArcIMS is a server-based product that provides a scalable framework for distributing GIS services and data over the Web (http://www.esri.com/software/arcgis/arcims/about/overview.html)
BATS	Bermuda Atlantic Time-Series station
BODC	British Oceanographic Data Centre
CCHDO	CLIVAR and Carbon Hydrographic Data Office
CIESIN	Center for International Earth Science Information Network
CLIVAR	Climate Variability and Prediction project
CTD	conductivity-temperature-density measurement package
CTDO	conductivity-temperature-density-oxygen measurement package
DAAC	Distributed Active Archive Centre
DAC	Data Assembly Centre
Data orphans	Data either unforeseen due to development of new technologies or lack of mandate, awareness, and/or capability at a DAC
DIF	Directory Interchange Format
DIU	Data Information Unit
DOI	Digital Object Identifier
DLO	Data Liaison Officer
DMC	Data Management Committee
DMTT	Data Management Task Team (JGOFS)
EEZ	Exclusive Economic Zone

ENTRI	Environmental Treaties and Research Indicators
ESI Viewer	Environmental Sustainability Index Viewer
GEOHAB	Global Ecology and Oceanography of Harmful Algal Blooms programme
GISS Crop-Climat	Goddard Institute for Space Studies Crop-Climat Study
GLOBEC	Global Ocean Ecosystem Dynamics project
GCMD	Global Change Master Directory
HOT	Hawaii Ocean Time-series
HPLC	high-performance liquid chromatography
IAPSO	International Association for the Physical Sciences of the Oceans
ICES	International Council for the Exploration of the Sea
IGBP	International Geosphere-Biosphere Programme
IOC	Intergovernmental Oceanographic Commission
IMAGES	International Marine Aspects of Global Change project
IMBER	Integrated Marine Biogeochemistry and Ecosystem Research project
Instantiation	the conversion of a virtual object (such as a dataset description) into a concrete object (such as a data file holding the dataset)
IODE	Intergovernmental Oceanographic Data and Information Exchange
IPO	International Project Office
JGOFS	Joint Global Ocean Flux Study
KNOT	Kyodo North Pacific Ocean Time-series station
LDEO	Lamont-Doherty Earth Observatory
LOICZ	Land-Ocean Interactions in the Coastal Zone project
MAST	Marine Science and Technology
NERC	Natural Environment Research Council (UK)
NODC	National Oceanographic Data Centre
OAIS	Open Archival Information System
OBIS	Ocean Biogeographical Information System
ODV	Ocean Data View
OMEX	Ocean Margin Exchange project
PetDB	Petrological Database of the Ocean Floor
PI	Principal Investigator
PROOF	PROcessus biogeochimiques dans l'Océan et Flux" = Biogeochemical processes in the Ocean and Fluxes
PROVESH	Processes of Vertical Exchange in Shelf Seas (MAST)
PSDS	Process Study Data Specialist

SAN	Storage Area Network
SCOR	Scientific Committee on Oceanic Research
SDS	Shipboard Data Specialist
SedDB	Integrated Data Management for Sediment Geochemistry
Semantic data standards	standards that unify the way in which data are described by metadata, such as controlled vocabularies (lists of approved words and their definitions) and content standards
SESAR	Solid Earth Sample Registry
SIO	Scripps Institution of Oceanography
SSC	Scientific Steering Committee
SME	small to medium-sized commercial enterprise
SOLAS	Surface Ocean – Lower Atmosphere Study
Soft-typed elements	items of metadata that are loosely defined by the metadata schema. For example, <temperature>25.0</temperature> is hard-typed, but <data parameter="temperature">25.0</data> is soft-typed
Syntactic data standards	standards that unify the way that data are physically encoded in a file (for example, CSV or NetCDF)
TEIs	trace elements and isotopes
US-Mexico	DDViewer: U.S-Mexico Demographic Data Viewer
WDC	World Data Centre
WDC-MARE	World Data Centre for Marine Environmental Sciences
WOCE	World Ocean Circulation Experiment