CDIAC - WHPO/CCHDO Data Management Plan for CTD/Hydrographic/CO_2/Tracer Data for the Global Ocean Carbon and Repeat Hydrography Program

Alex Kozyr
Carbon Dioxide Information Analysis Center
Environmental Sciences Division
Oak Ridge National Laboratory
U.S. Department of Energy
Building 1509, Mail Stop 6335
Oak Ridge, TN 37831-6335 U.S.A.

tel 1-865-576-8449
fax 1-865-574-2232

kozyra@ornl.gov
http://cdiac.esd.ornl.gov/oceans/home.html

James H. Swift
WOCE Hydrographic Program Office
(CLIVAR and Carbon Hydrographic Data Office)
UCSD Scripps Institution of Oceanography
9500 Gilman Dr., Mail Code 0214
La Jolla, CA 92093-0214 USA

tel 858-534-3387
fax 858-534-7383

jswift@ucsd.edu
http://whpo.ucsd.edu

Summary

Effective management and archival of data is a fundamental requirement for successful scientific research endeavors, and future oceanographic research depends on the availability and clarity of existing data. Two data offices in the US deal with reference-quality global ocean CTD, water sample, and underway data, one (CDIAC) specializing in discrete CO_2 and underway surface data, and the other (WHPO/CCHDO) specializing in CTD, hydrographic, and tracer data.

Since 1993 CDIAC has been serving the ocean scientific community as the central repository for the carbon dioxide data measured on the WOCE/JGOFS cruises. CDIAC receives WOCE hydrographic and tracer data from the WHPO. Thus all US and most foreign WOCE hydrographic, chemical and carbon data are available now through the CDIAC WWW. Most of the data at CDIAC are available as published and electronic Numeric Data Packages (NDPs). The CDIAC_WOCE Ocean Data View (ODV) Collection that includes all WOCE sections with

CO_2 measurements as well as hydrographic and nutrient measurements is now available through CDIAC WWW. CDIAC communicates frequently with the scientific measurement groups and individual PIs. This has helped CDIAC build the largest atmospheric and oceanic carbon data sets in the world, with the highest quality data. As the new carbon data measurements will be measured by the US and foreign measurement groups on the repeat hydrographic sections, CDIAC is ready to continue its support to the WHPO/CCHDO in CO_2 data processing and archival.

The fundamental role of the WOCE Hydrographic Program Office (WHPO) [soon to be known also as the CLIVAR and Carbon Hydrographic Data Office (CCHDO] at the UCSD Scripps Institution of Oceanography is to see that WOCE Hydrographic Program data, CLIVAR repeat hydrography data, global ocean carbon hydrographic data, and other similar CTD/hydrographic data and their associated documentation are prepared and made available for both immediate use and a long service life. The CTD, hydrographic, and tracer data used in large scale ocean circulation studies are brought together, verified, corrected for content and format errors, assembled with relevant documentation, and carefully prepared for dissemination and archive. In addition the WHPO works to promote appropriate methodology, applicable community standards, communications, and data compatibility. The WHPO supported these important functions for the WOCE Hydrographic Program data from 1997-present, and has been invited to continue these functions for CLIVAR hydrography, global ocean carbon hydrography, and similar programs which make use of high quality ocean profile data. Data of the type dealt with by the WHPO are created by >100 data originators worldwide, sometimes 5 or more contributing to one file. All data users must cope with the temporal-, content-, and format-related file diversity these different originators engender. It is the enormous advantage of bringing data sets together to a common content and readability standard that remains the key function of the WHPO, with a strong additional advantage that the documentation associated with the data are collected, reorganized to a common standard (where possible), and preserved with the data. Although the data office disseminates data via the internet and on CD-ROMs and data DVDs, it provides its total public holdings, including documentation, to NODC/WDC-A for archive and further distribution.

CDIAC and the WHPO cooperate closely: CDIAC receives many CO_2-related data files directly, and also some from the WHPO. CDIAC carries out all data management functions for CO_2-related data and the WHPO handles these functions for the CTD, hydrographic, and tracer data. The WHPO merges into its data files the latest versions of the CO_2-related data as received from CDIAC. CDIAC uses the latest versions of the hydrographic data in its files. The WHPO is the primary provider of the data to NODC/WDC-A. Both facilities distribute data in formats agreed to be their user communities.

US Global Ocean Carbon and Repeat Hydrography Program

A systematic and global re-occupation of select hydrographic sections is underway to quantify changes in storage and transport of heat, fresh water, carbon dioxide (CO_2) and related parameters. By integrating the scientific needs of the carbon and hydrography/tracer communities, major synergies and cost savings are being achieved. The philosophy is that in

addition to efficiency, a coordinated approach will produce scientific advances that exceed those of having individual carbon and hydrographic/tracer programs. These advances will contribute to the following overlapping scientific objectives:

- Data for Model Calibration and Validation
- Carbon System Studies
- Heat and Freshwater Storage and Flux Studies
- Deep and Shallow Water Mass and Ventilation Studies
- Calibration of Autonomous Sensors

Earlier programs [e.g., World Ocean Circulation Experiment (WOCE)/Joint Global Ocean Flux Survey (JGOFS) survey during the 1990s] have provided a full depth data set against which to measure future changes, and shown where atmospheric constituents are entering the oceans. The proposed measurements will reveal much about internal pathways and changing patterns. They will serve as a baseline to assess changes in the ocean's biogeochemical cycle in response to natural and/or man-induced activity. Global warming-induced changes in the ocean's transport of heat and freshwater, which could affect the circulation by decreasing or shutting down the thermohaline overturning, can be followed through long-term measurements. Below the level of the Argo array, repeat hydrography is the only global method capable of observing these long-term trends in the ocean. The program will provide data for sensor calibration (e.g., www.argo.ucsd.edu), and to support continuing model development that will lead to improved forecasting skill for oceans and global climate.

The US contribution to the global program will maintain decadal time-scale sampling of ocean transports and inventories of climatically significant parameters. The sequence and timing for the proposed sections takes into consideration the program objectives, providing global coverage, and anticipated resources. Also considered is the timing of national and international programs. In addition, the proposed sections are selected so that there is roughly a decade between them and the WOCE/JGOFS occupation.

The US work will be based on the success and experience that were built during the WOCE/JGOFS surveys. The most important approach for the new measurements will be consistency and standardization of techniques. It is assumed that measurement groups will perform 2-3 repeat sections per year, and collected data will have gone through their preliminary or primary research stage at the responsible measurement group.

CDIAC

Discrete CO_2 -related Data Collection and Submission

The CO_2 data will be available to the public within 6-12 months through the WWW with a preliminary tag. Once the Primary Research Data Management work is accomplished the CO_2 data will receive a Secondary Research status and will be moved to CDIAC. What format (e.g., ASCII flat or comma-separated) the measurement groups chose to submit data to CDIAC is not critically important, but consistent file formats, parameters, and units are essential.

QA/QC

It is assumed that the discrete CO_2 data will have gone through its preliminary (primary) stage at the institutions responsible for discrete measurements. During this time (up to 6 months?) the data should be carefully examined by the responsible PI for possible obvious outliers and internal consistency, the corrections for post cruise calibrations should be applied, and the quality flags for the carbon data should be assigned. Also, decisions on possible adjustments based on shipboard CRM analyses should be made.

As soon as the discrete CO_2 data from the repeat sections are transferred to CDIAC as Secondary Research Data, CDIAC will release the data the CLIVAR/WHPO office and to public via the WWW Live Access Server (LAS) with a Secondary Research Data tag and perform the basic QA/QC

Received carbon-related data will be merged (if necessary) with the final hydrographic measurements (if available) and the file will be put in the uniform format (old WHPO format, CSV format, other?). If the hydrographic and chemical data are not available at the time of receiving the carbon measurements, CDIAC will contact the CLIVAR/WHPO office and work together in preparation of the final data set.

If problems with data are found during the QA-QC, CDIAC will contact the responsible PI(s) and CLIVAR/WHPO. PI, CDIAC, and CLIVAR/WHPO will work together to resolve all problems in order to upgrade the data to Archival (Final) data status.

Metadata Requirements

Each data set sent to CDIAC should be accompanied by the required metadata information (e.g., investigator names, carbon measurement method and instrumentation description, measurement precision and accuracy estimates). The metadata should be sent to CDIAC in a consistent format as a separate file. CDIAC will be responsible for the integration of metadata into the appropriate place in the LAS along with measured data.

Delivery of Carbon-related Data to Researchers and the General Public

The repeat section carbon-related and hydrographic data will be made available to users through the WWW-based LAS server, ODV format, and as a simple ASCII formatted files through CDIAC's public FTP area.

CDIAC has implemented the LAS system in 2001 and will establish the WWW CDIAC/LAS server for the repeat section measurement. Users will be able to access these data and metadata through CDIAC/LAS server and other LAS sister-servers established at AOML and PMEL and other groups. For users who cannot access the LAS WWW servers, CDIAC will offer the data as ASCII data files (fixed format, comma-separated format) through the CDIAC FTP area. CDIAC will continue to publish NDPs for each repeat section as soon as the carbon-related and hydrographic data have Final status. Also, CDIAC will offer to users the ODV Collection of the

repeat sections data through the WWW.

Permanent Data Archival

As soon as the discrete carbon data measured along the repeat sections reach Archival (Final) status, CDIAC will send all data and metadata files to CLIVAR/WHPO for the final merging with the hydrographic and other data. CLIVAR/WHPO will provide their routine QA/QC on hydrographic and other data and will contact CDIAC in case problems are found. It is assumed that after the data receive Archival (Final) status there will be no further corrections applied to the measurements by responsible PI(s); however, if the PI(s) makes a decision to apply corrections to the final data he or she will notify CDIAC about the action. CDIAC will then notify CLIVAR/WHPO about corrections.

CDIAC will play a role in the long-term archive for all repeat section measurement data sets. CDIAC will store all repeat discrete measurements, along with all companion documentation and metadata, on the CDIAC Computing System.

WHPO/CCHDO

The most important WHPO functions are:

(1) locating data and arranging for data and documentation transfer to the data office,
(2) checking all data and headers for errors and correcting those errors,
(3) merging bottle data parameters from disparate sources,
(4) moving the data into tight agreement with well-specified community data formats,
(5) bringing together, organizing, and preserving the information about the data necessary to understand and use them, and
(6) providing for widespread distribution of the data and archive of the data.

In addition, it is of great value to have good communications between the data office and the data providers beginning ahead of cruises to provide resources and advice which will help those at sea to achieve the data quality and reporting standards that will best meet long term community needs.

Data acquisition and assembly: The data office acquires data and documentation in activist mode, i.e. soliciting individual investigators and data providers (not only Chief Scientists) by mail, email, fax, telephone, inquiries and reminders to investigators, institutions and national committees. The Director frequently discusses data matters with individual investigators. Every effort is made to ascertain difficulties and to find a means to successfully acquire data.

Data are accepted in any readable format but preference is given to the formats supported by the office. Preliminary data and documentation are accepted, in which case the data office works with investigators to arrive at a mutually-acceptable timeline for completion. The data office maintains a data catalog and tracks and documents data modifications as well as ancillary information such as measurement and submittal dates, proprietary status, and so forth for each

data subset.

The office acts as a data assembly center. For example, bottle data parameters are typically received from several investigators. Nearly every imaginable problem seems to arise. Because contacts with the data originators are often required, and because the office, in the interests of accuracy, cannot take some of the arbitrary data actions data users often do, it can take months to unambiguously merge some data (though many other data mergers are simple and quick).

Quality control: The data office uses a hierarchy of data quality control procedures to ensure the highest possible standard of data accuracy and utility. The process includes providing example data and documentation files, validating data for content and format, and preparation of a corrected-for-readability version of the data file.

More often than not, data are received in less than desirable condition. Problems include conflicting header and sample information, missing values, incorrect values, wrong units, lack of information about the files themselves, version control problems, data in individual station files or continuous strings, data not in ASCII, etc.

All too often a data originator is unreachable or uncooperative. With this in mind it is no surprise that repairing data problems is truly a massive undertaking. Though the office has worked hard to streamline this, for some cruises there is a great deal of work required.

The files for many cruises are dynamic and multi-dimensional, with the office holding several disparate copies. For example during the WHP, the data did not follow the acquisition, processing, submittal, DQE, adjustment, and archive sequence envisioned by the WHP planners. Rather there was a semi-chaotic process by which preliminary files and subsequent updates were received in near random order (for the various parameters) over a period of years, with some new files referenced to out of date versions of old companion files. Common issues include changed bottle or sample numbers, evidence of investigators working with data within their laboratories in their traditional (non-WOCE) format, tracer samples reported for stations, casts, and/or bottles not in the S/O2/nutrient file, and indexing by depth/pressure instead of station, cast, and bottle/sample number. It can be a major exercise to make changes which sweep across a swath of files of various vintages for the different parameters.

The WHPO, unlike a data user, cannot simply fill in values to make a file import easily. WHP file inconsistencies presented a continuing problem for WHP data users that was not resolved until the WHP-Exchange formats were created. With very few (if any) exceptions, if one can read one WHP-Exchange data file, one can read any other WHP-Exchange data file. This eliminated the disturbing need for some data groups to devote significant amounts of staff time to importing WHP data.

During WOCE there was a subsequent data quality control process involving intensive Data Quality Examination ('DQE') carried out by community data experts. Based on the value returned for effort spent, and in particular the fact that data *values* were almost never changed during the DQE process, the most valuable quality control activities were the often-laborious processes of cleaning the data files, parameter by parameter, for errors and improved adherence

to specified community formats. The WHPO does not plan to continue external DQE in the post-WOCE era but does intend to carefully validate data files. Data for some parameters (e.g., CO_2-related parameters and CFCs) occasionally undergo external quality review by groups outside the WHPO (e.g., CDIAC for CO_2-related parameters). The WHPO assimilates the results of these DQE activities into its data files and documentation.

Data access: Access to data held by the office is provided primarily by on-line service and secondarily by data CD-ROMs, which are essentially copies of the on-line site.

The files consist of some to all of the following file types (primary file types in bold):

.sum	WHP cruise/station/cast summary information essential for data file interpretation
.hyd	_hy1 rewritten in original WOCE format (entire cruise)
.ctd	_ct1 rewritten in original WOCE format (individual files in gzip'd directory)
.lvs	WHP large-volume sample data file [no longer in use]
.doc	WHP documentation; in gzip'd directory containing ASCII and .pdf files
_hy1.csv	WHP bottle data file (entire cruise) in WHP-Exchange format
_ct1.csv	WHP CTD data file (individual files in gzip'd directory) in WHP-Exchange format
_hy1.nc	_hy1 rewritten in netCDF format (entire cruise)
_ct1.nc	_ct1 rewritten in netCDF format (individual files in gzip'd directory)

Data version control and data originator citation: The office has long kept track of and made available data history and data originator information for every data set. But the office is now preparing to go beyond this by embedding key file version, investigator, and citation information in the 'comment' lines in new WHP-Exchange data files for the CLIVAR and global ocean carbon programs. This will list all previous versions of that file, all data originators, plus the American Geophysical Union requirements for citation of data, as in this made-up example for the header of a WHP-Exchange bottle data file:

```
BOTTLE,20000814WHPSIOJJW
#BOTTLE,20000612WHPSIODAM
#BOTTLE,19991103WHPSIODBK
#code : jjward hyd_to_exchange.pl V1.0
#original HYD file: p23chy.txt   Thu Aug 10 14:31:46 1999
#original SUM file: p23csu.txt   Thu Aug 10 14:31:46 1999
#These data were provided by
#parameter/program   name                email
#Chief Scientist     James H. Swift      jswift@ucsd.edu
#CTDO/S/O2/nutrients James H. Swift      jswift@ucsd.edu
#CFC                 William M. Smethie  bsmeth@ldeo.columbia.edu
#He/Tr               Peter Schlosser     peters@ldeo.columbia.edu
#14C/13C             Robert Key          key@geo.princeton.edu
#CO2                 Richard Feely       feely@pmel.noaa.gov
#It is required that those who use these data before 01 January 2005
#inform the data provider(s).  Before this date the data are preliminary.
#The data providers must be acknowledged in any presentation, publication,
#or proposal which refers to or uses these data.  Data use shall be cited
#as: "data provider(s), cruise name or cruise ID, data file name(s),
#CLIVAR and Carbon Hydrographic Data Office, La Jolla, CA, USA, and
#data file date."
```

Data archive: The office provides all publicly-available data and documentation to the archive at NODC/WDC-A.

Documentation: The office prepares as complete a documentation set as feasible for each cruise, including cruise reports, final reports and amendments received from the Chief Scientist and other participating investigators, and summaries of in-office data notes and file information. The files are prepared in two formats: The primary copy is a pdf (portable document file) document containing all of the text, tables, and graphics. A second version is ASCII plain text.

Investigator support: The office maintains on-line information about each cruise, including (among much else) investigator contact information. Maps and search capabilities guide investigators to data of interest. Additionally the office maintains an on-line library of electronic documentation of cruise reports and other documents.

Oversight: The Director is responsible to community oversight groups and NSF. The Director prepares and discusses with the oversight bodies and NSF an annual assessment and work plan, plus the Director maintains frequent contact with the appropriate international and US project offices.

Data CD-ROMs/DVDs The pace of updates warrants publication of annual updates on CD-ROM. These serve data users without internet connection (e.g., at sea), slow internet connections, or who need a large fraction of the data collection at hand. The latest full copy requires four CD-ROMs and so a data DVD was published.