POGO: Data Issues and Proposed Way Forward

Background:

Without a doubt, management of, and access to data are major problems that face oceanographers today. The success of any observational scheme depends heavily on timely distribution of the data to the user community. This requires a concerted effort, and POGO should do whatever possible to find solutions to problems in this area.

Several international agencies and organisations (for example: JGOFS, OBIS, CLIVAR, JCOMM, GOOS, PICES, IODE, IOC) are aware of the issues that face oceanographic data management and distribution, and have pooled expertise and resources into solving the problems. Whatever POGO does has to be well co-ordinated with these other efforts, to avoid duplication.

There is no expertise on the subject matter at the POGO Secretariat itself. Help from experts is needed, to make effective progress.

- At POGO 3 (November 2001), it was therefore resolved to recruit experts from the data management arena from POGO institutions to help the POGO Secretariat. It was suggested that this could be achieved through the establishment of a small working group of two or three experts with experience in handling and distribution of multi-disciplinary data. Dr. Savi Narayanan, Co-President, JCOMM, who was present at the meeting, agreed to be a member of this expert committee. It was also resolved to approach Dr. Roy Lowry, Proudman Oceangraphic Laboratory, to be a member. Dr. Lowry has since agreed to be a member of this committee. The third, non-expert member of the ad-hoc committee is to be Shubha Sathyendranath, Executive Director of POGO.

- At POGO 3, it was also recognised that a workshop being organised by Dr. Neville Smith could be an excellent forum to forge a common way forward, and to identify the role that POGO could play, in co-ordination with the other players in the field.

In April 2002, Savi Narayanan and Shubha Sathyendranath met at a JCOMM meeting. They used this opportunity to discuss data issues between themselves, and to develop some more detailed plans for action. The thoughts that were developed at this meeting are summarised below.

The Major Data Issues in Oceanography:

The major issues are identified as follows:

1. Not all data collected in oceanographic institutions are stored in electronic databases, making it difficult to safeguard, use and exchange data. This may
be related to the local, institutional approach to the management of scientific
data collected by the staff.

2. Mechanisms are not in place everywhere to facilitate timely dissemination of
data to the user community.

3. Exchange of data between oceanographic institutions is not always easy.

4. There is insufficient awareness of efforts currently underway in various
organizations and programmes to address issues related to data management,
and data dissemination. This in turn makes it difficult to contribute to, or
benefit from, such endeavours.

5. Management and distribution of data on non-physical variables and
descriptive information are particularly challenging.

**Brief to the POGO ad-hoc committee on data issues:**

- Catalogue the data management practices in member institutions.

- Identify impediments to enhancing data archive, data exchange and
dissemination. Suggest solutions within reach of POGO, when appropriate.

- Report to POGO on efforts related to data management and data dissemination
that are already underway in various organizations and within international
research programmes. If appropriate, identify what POGO members could do to
enhance these efforts.

- Identify issues related to archival and distribution of non-physical variables and
information. Inform POGO members of efforts that are already underway to
address these issues. Recommend appropriate action.

**Issue 1. Data Archive**

As vast amounts of data are collected by oceanographers, commensurate effort has to be
invested in archiving and managing the data. Otherwise there is grave danger that the
data may be lost, as researchers and technicians move away or retire. This problem is
perhaps more acute in the case of non-physical variables and descriptive information. As
technologies and planning advance, various schemes are emerging to address these
issues. It would be premature now to advocate for a particular system or a specific
approach. The important thing at the moment is to ensure that the information is archived
in appropriate databases, which would facilitate translation into other systems as and
when necessary. Oceanographic institutions have varying levels of expertise and
experience in dealing with these problems. Institutions can share information and expertise on how to address this issue.

**Action:** POGO Secretariat to query POGO institutions and other participating institutions on their data management practices and major concerns with respect to data archive and data management. Can they suggest ways in which institutions could help each other to address these issues? What is the one, single most important thing to be done, feasible in the near term, that would facilitate data exchange amongst the ocean observing community, and/or the POGO community? What steps should POGO take to advance the data exchange agenda? Results of enquiry to be presented at POGO 4.

**Issue 2. Data Dissemination to the User Community**

The dissemination of data to the user community carries with it the need to make the data available in a timely fashion. Often, this means dissemination of data in real time or near-real time. As we move towards operational oceanography and sustained measurements, the success of our efforts will be evaluated on the basis of how useful the enhanced measurements have been to improve the delivery of planned products. This cannot be achieved without addressing the issue of timely delivery of data to user agencies.

**Question:** Are oceanographic institutions, especially those with a research mission, able to provide data in real time or near-real time to operational agencies? If not, what are the impediments and how can they be addressed?

**Action:** Take a specific case and see if POGO can do anything to improve the performance in that area. JCOMM has identified the need to ensure the global availability of XBT/CTD (and other) data, both in real time and in delayed mode. In this context, it would be worthwhile to encourage transmission of such data from research vessels to the appropriate data centres in near-real time (within one month). The one-month time limit decreases the demands on resources required for transmission of data in real time by GTS from research vessels. Some institutions have addressed this issue in a cost-effective manner by using the ftp method to transmit data to data centres. These centres then take the responsibility of making the data available on GTS. POGO Secretariat to query POGO members and other participating institutions on their policy with respect to collection and delivery of XBT/CTD data. What can be done to enhance the performance? The results to be presented at POGO 4.

**Issue 3: Exchange of data between institutions and the research community**

There is a need to enhance access to data for use by the research community. This has two elements: (i) difficulties with electronic access; and (ii) difficulties with access to information on what is available at various sources.
With respect to the problem of electronic access, POGO already has, under the leadership of Dr. Robert Weller, initiated a data exchange pilot project that is based on the Distributed Oceanographic Data System (DODS). The effort can be continued, if there is sufficient interest among POGO member institutions.

With respect to the problem of advertising available data, the possibility exists of using the POGO web site to advertise information on data available at various sources.

**Action:** POGO Secretariat to begin advertising data sources on the POGO web site. Roy Lowry and Savi Narayanan to pass on any comments or suggestions on this initiative to Shubha Sathyendranath. Shubha to discuss progress in implementation with Savi Narayanan and Roy Lowry.

**Issue 4. Linking to Existing Efforts**

Dr. Neville Smith is participating in a workshop on data issues in conjunction with the PICES annual meeting in October 2002. This workshop is designed to take stock of existing efforts and chart the way forward.

**Action:** POGO Secretariat has requested Dr. Smith to report at POGO 4 on the findings of the workshop, and recommend possible action that POGO could take to enhance these efforts. Dr. Smith has agreed to make a report at POGO 4.

**Issue 5. Providing Impetus to Management and Distribution of Non-Physical Data**

The archiving, management and distribution of physical data function at a much more advanced level than is the case for non-physical variables, parameters and descriptive information. As POGO moves towards incorporation of biological measurements into its observation plan, it is important to ensure that the data management and dissemination keep pace with the observations of non-physical variables. A strategy has to be developed in this regard.

**Action:** Roy Lowry to recommend possible actions that can be taken in this area, either through the individual institutions, or through the POGO Secretariat. The recommendations to take into account the Actions recommended in the POGO Biology Report. These recommendations, after discussion within the POGO Ad Hoc Committee on Data Issues, to be presented at POGO 4 in Hobart, in January 2003.
Responses to POGO Questionnaire on Data Issues

1. Institutional Data Policy:

CSIRO (Australia): has a data policy which is being reviewed.

IAO (Argentina): The institute has a data base of charts, aerial photos, etc. Because every researcher gets his/her own funds to perform their research, the institute does not have a data base of the measurements that the researchers have. The researchers take care of their own data.

IFREMER (France): The present data policy for physical oceanography at IFREMER, which also operates the French National Data Centre (SISMER) is:

1) to disseminate the raw CTD and XBT data (and possibly other parameters in the future) in real time, by automatic procedures on board the main Research vessels, or by deploying drifting profilers (Argo/Coriolis project); these data are put on GTS.

2) to disseminate the validated scientific data typically 2 years after the cruise implementation; inter-organisational panels encourage to take action to reduce this delay. The data are compiled from data requests to the responsible laboratories, but as the priority is to safeguard the data, if the scientists request longer confidentiality time, it is accepted for the moment.

3) more and more data are collected in the frame of cooperative projects, especially EU projects for which a data management plan is in general requested. The data management is made with standards developed by partners data centres and based on the IOC and ICES guidelines when available.

4) a new action is in preparation on a "pan-European scale" (European and Mediterranean countries) to enhance these standards, network the data centres and facilitate the on line access to "integrated data sets from semi-distributed sources" (SeaDataNet network). BODC and IFREMER/SISMER will be among the key partners.

This action can meet several of the data management issues raised by N. Smith. One of the key issue, which is not yet critical for the operational oceanography, is the compatibility of standards which have been developed in parallel by different sub-groups, for many parameters: formats, meta-data, quality control procedures.

IMR (Norway): has an integrated data base primarily designed for internal use. As a rule, all quality controlled data are free to external users. However, restrictions may apply for a limited time on individual data sets. It is the Institute’s intention to make
subsets of the data available on line. Free data are not available on line, but will be available on request.

KORDI (South Korea): The oceanographic data collected in KORDI is managed according to the following 3 steps.  
Step 1: The data management section tries to collect all Metadata of each observation as soon as possible. Database and internet service system is established and operated to provide information about data production.  
Step 2: The data files presented by each research group is linked to Metadata DB and provided to all researchers of KORDI through internet.  
Step 3: Some data like temperature, salinity, satellite data, which every researcher wants to use immediately, is archived in a separate database system and disseminated to user community.  

PML: is in the process of developing a policy.

2. Major concerns with respect to data archive, management and distribution:

Combined responses from all institutions.

a) Scientists do not generally build the cost of data management into their research funding proposals so it is hard to get projects to allocate resources to manage acquired data as a Divisional asset and in accordance with Divisional guidelines.

b) There may be an Institutional Data Policy but most projects do not comply with the Policy.

c) There is often no consistent approach within an institute on data access, data sharing and public data release (this may be true even when guidelines exist). The two extremes are "all data and products should be released immediately into the public arena" on the one hand, and "most data release should be undertaken through a case by case examination and caveats might be placed on access to some data" on the other hand.

d) Data re-use is restricted by poor data management practices. This eventually costs the Institutes money.

e) Intellectual property rights. Who actually “owns” the data?

f) How do we determine the value of data? Do they have any value at all on their own?

g) It is not possible to find any quantitative data on cost/benefit analyses associated with the benefits of good data management practices.
h) New types of instrumentation now generate vast quantities of data that are poorly mined and utilised because little investment is put into post capture processing and analysis. The media that the data is captured on are not properly archived so the ability to look at the data as new techniques and tools are developed diminishes quickly with time.

i) Insufficient software to handle data bases. Need more courses in data management.

j) Quality control of data, especially chemical and biological data.

k) Management of data produced by newly-introduced instrument.

3. **How can Institutes help each other?**

**CSIRO:** Some of the deficiencies expressed above are issues often partly driven by national government policies on things like cost recovery, intellectual property rights, competition for research funds and research funding mechanisms. However, many issues are also the result of specific institutional cultural practices. Researchers are not generally intrinsically (professionally) rewarded for sharing data, in fact they can benefit by not doing so. Many papers can be milked from one comprehensive, historical time-series data set.

Perhaps institutions can examine ways of rewarding researchers that adhere to good data management practices. Sharing best practice examples, that clearly show in qualitative terms, how good data management practices benefit research projects and institutions, always helps. Raising the profile of Information Management within institutions by employing senior IM professionals as part of institutional management teams will demonstrate and drive commitment from the top and lead the cultural change required. A quantitative study could be undertaken to examine, in dollar terms, the cost/benefit of good data management practices.

**IAO:** People do not want to share data. Therefore sometimes it is hard to get data from them. May be courses or experiences in other countries may help to change their concept.

**KORDI:** A cyber community can help exchange of technical information about data management.

4. **Potential Steps that POGO can take:**

a) Make sure it is linked in with what is happening on the international scene with respect to information management and is cognisant of emerging policies and issues.

b) Sponsor a quantitative cost/benefit analysis study as described above.
c) Follow the path POGO seems to be taking, consult institutions widely, look for achievable actions that might address the issues at the grass roots level (within the institutions themselves) and promote a program of work to address the issues. Some cross-institution collaborative work would be all the better.

d) Some scientists feel that they are being used if they give their data. Perhaps with conferences explaining the importance of sharing data, international projects, etc, it may be able to produce a change in their concept about data management.

e) Collecting and archiving all data in a data base system is impossible, because the number of data items is increasing and volumes of oceanographic data is enlarging. In the first place, establishing distributed Metadata data base network using XML technology (marine XML of IODE) is required to share information on where the useful data is.

f) Workshop or training programme about technology of data management and exchange can decrease the technological gap among the institutions.

g) The key things are to maximise the exchange of data, and to enhance the existing mechanisms of ICSU (WDCs) and IOC.

5. **Response to JCOMM Request:**

**Kim Finney, CSIRO:** XBT data (through ship of opportunity programs) goes directly onto the GTS and I believe new software to be installed on our National Facility Research Vessel will transmit low-resolution data in near real time. All routinely collected physical data sets are transmitted in slow time to the US NODC. Copies are provided to the Australian national Centre upon request.

What can be done to enhance the performance? Develop the appropriate software and mechanisms to make real-time transmission an easy, routine automated task. Make sure that the appropriate policies are in place to ensure that Principal Investigators are obligated to transmit data in real-time. Not all data collected on our National Facility falls under the same data release policy. Nationally we need to review how we operate our National Data Centre. I understand we may go to a distributed model.

**Nick Owens, PML:** Typically, PML follows NERC Guidelines on XBT/CTD data, with these data going eventually to BODC; the data are then subject to NERC’s Data Policy.

**Oeyvin Strand, IMR:** IMR hosts the national data center of Norway, and regularly delivers CTD, and other data, to ICES and the WDC system. As a main rule, data are withheld from external distribution until quality controlled. This policy will often impose a delay on international distribution of data sets.
Cintia Piccolo, IAO: We contribute to our national data centre, but with a time lag of many years. Some researchers complain about not getting the data from the National data centres, and so are reluctant to give them the data. An example is the tidal data. The Argentinian Oceanographic Data Centre has data from some coastal zones, but it is hard to get the data.

Sung-Dae Kim, KORDI: The database of CTD data is operated and an internet system provides it for user community. Recently, real time data are collected at some coastal stations and serviced through internet.

Advertising Data Sources:

CSIRO: Please see CMR's data holdings via our metadata database known as MarLIN (at http://www.marine.csiro.au/marlin/) also our new data warehouse and data delivery tool called Trawler (at http://www.marine.csiro.au/warehouse/jsp/loginpage.jsp). The trawler is a new product and is being trialled for data discovery and data dissemination within the Division. It does allow the public access but not all data in the Trawler is available to the public. To see some public data look for categories of data related to the research vessel RV Franklin. These data sets can be downloaded by the public.

IFREMER: Our data portal to the different IFREMER databases is presently: www.ifremer.fr/sismer
A new version is under development with a more integrated presentation and an improved on line acces to data; it will be opened by the end of 2002.

KORDI: The realtime oceanographic data system is developed and operated by the KORDI, carrying out realtime data acquisition, processing and data service of 9 oceanographic and meteorological data at 12 observation stations. Only Korean version was available in August, but the english version was expected to be ready in October 2002. The address is http://realtime.kordi.re.kr

Summary

The general feeling is that what is desired of POGO is not new initiatives at data exchange, but promotion of sound data archival policy within oceanographic institutions. Some specific suggestions follow:

1. **Contribute to Training.** There is expertise in many POGO institutions on data archival, management and distribution. Experts from these institutions can contribute to training of new experts in other institutions, either through training courses (Ocean Teacher), or through the POGO-IOC_SCOR Fellowship Programme.
2. **Identify Data Issues as a priority area** for the POGO-IOC-SCOR Fellowship Programme.

3. **Institutions have to invest more in data archival.** The first step towards good data management has to be timely archival of scientific data. The access to data may be restricted as appropriate, to protect, or limit access to data according to accepted terms.

4. **Scientists to be encouraged** to provide data to institutional data bases. One way to do this is of course by educating the scientists of the advantages of sound data policy. Another may be through institutional policies that consider timely contribution of data to institutional data bases as part of the criteria for performance evaluation.

5. **Twin Institutions.** It is clear from the responses received that there is a huge disparity between oceanographic institutions in capacity for data management and distribution. If an advanced institute can “adopt” one of the lesser developed institutions and help them to improve their capacity, it would be a tremendous step forward.

6. **Operational Requirements.** From the responses received to the POGO questionnaire, it is not evident that all oceanographic institutions have a good record of providing data in a timely fashion to meet operational requirements. In some instances, the poor response may be linked to national data policies. When this is not the case, the POGO Directors can play a role in facilitating and encouraging transmission of data to meet operational needs.

7. **Use POGO Web Site.** POGO web can be used to advertise existing sources of data. POGO Directors to encourage their data managers to bring data availability to the attention of the POGO Secretariat.

**POGO Contacts for Data Issues**

| BIO       | Doug Gregory       | gregoryd@dfo-mpo.gc.ca |
| IAO       | Cintia Piccolo     | ofpiccol@criba.edu.ar |
| IFREMER   | Catherine Maillard | Catherine.Maillard@ifremer.fr |
| IMR       | Oeyvin Strand      | nmd@imr.no            |
| KORDI     | Sung-Dae Kim       | sdkim@kordi.re.kr     |
| PML       | Nick Owens         | njpo@mail.pml.ac.uk   |
| SOC       | Lesley Rickards    |                      |