Data Centre Managers Meeting – Part 1
12-13 May 2010

Introduction

E.W. Hare* & E.J. Carty**

*Environment Canada

** Software Solutions

Rationale

Action items born out of the 2007 Integrated Global Atmospheric Chemistry Obs. (IGACO) meeting in Dübendorf, Switzerland (next slide) - Data Consumers have a long wish list – the question is whether data centres can deliver?

Also, the WOUDC has been serving the ozone (and more recently the UV) scientific community for nearly 50 years. Is it time to set a new course?

Perhaps the best approach to resolving data centre issues is through knowledge sharing, best practices and lessons learned from the various “extended” communities.

Follow-up Issues – IGACO

Activity D1: Better data access and archiving,

Activity D7: Multiple versions of same data

Activity D9: Overview of existing data services

Activity D11: Identifying historical data

This meeting is in response to these bullets

Purpose

To discuss issues such as:

- Multiple data centres doing similar but not necessarily complete data archiving
- Level 0 “raw” data versus Level 1 “ data reports” versus Level 2+ custom “gridded” data
- Data “deployment”
  - Data push vs pull
  - Data files vs Databases
  - Binary vs ASCII
- “Long-term” archives versus snap shot databases OR project-based archives

Evolving influences

Since the Internet our role, responsibilities and community has changed (and broadened), perhaps beyond the initial "scope" and original mandate – which is to serve as a long-term data archive only – and perhaps the capacity to deal with all these elements.

For example, the centre now must:

- Provide aspect of community and media relations coordination (framing)
- Become data quality experts (not our initial purview)
- Manage larger volumes of data (increased a 1000 fold+)
- Be subject matter experts (in the case of the WOUDC – that is 7 areas of focus)

Present Status – A WOUDC Perspective

In General:

- The WOUDC is a "**long-term**" archive – with that responsibility comes many constraints – for one, Tradition
- It is not the WOUDC's purview to determine how data are used –similar to a library – we distribute the books (data) but the Intellectual Property (IP) (book content) is the authors

Two things become apparent:

1) **Recognition and citation of Intellectual Property by data consumers** (enhances viability of data producers and data centres)
2) **Enhanced use of Version and Data Sponsorship Statements by data producers** (enhances data quality and extensibility)

Some Problems & Issues identified at the WOUDC

- Level 1 data are always evolving
- Thus, translators are time consuming and inconsistent
- Inconsistent versioning and handling of revisions etc.
- No standard 'metadata', data model (consistent data dictionary)
- No standard 'data content', data model (consistent data dictionary)

A New Paradigm?

Challenges to Data Managers – Are we still relevant?

1) Is there a need / desire for WEB2 type re-development of WMO Data Centers?
2) Is this the right time for wholesale WMO Data Centres re-development?
3) How do we address the ongoing proliferation of data discovery portals?
4) Distributed databases? Data Centres API development?

Research vs Operational Needs

Data Centres must confront the difference between Research and Operational needs.

Research based data;
- seeks discovery and innovation
- accepts errors and learns from exceptions

When data are in the research domain the people "closest" to these data are usually the only ones in a position to produce these data products. The responsibility for data quality and level 1+ data products ultimately lies with the Data Owners.

Operational based data;
- produces known products from stable sources
- seeks to minimize errors and can fail due to exceptions

Once data become operational, these processes can be offloaded to Data Centres, who can archive level 0 (collection) data and provide standard methods (algorithms and processes) for producing a variety of level 1+ data products. In this model, level 1+ data quality can be the responsibility of the Scientific Community (implemented by the Data Centres).

In either scenario, Intellectual Property of the Data Owners must be preserved.

Different Views of Data

Sample Data Consumer needs;

- Modellers (forecasters) want data in near-real time – want data quickly, but will accept errors up to 10%
- Data Validators – less “urgent” (will wait up to several weeks) but want 3-5% data accuracy
- Trends and Assessment Scientists (patiently wait for several months) but want data to be 1-2% accuracy

Data quality, time frame, etc. are driven by the different needs. Often one need can oppose another.

Can we meet the needs of everyone, all the time?

Level 0, Primary ("raw") Data

- Current WMO-WDC framework is based on level 1+ data "publications"
- Different needs; "***one person's trash is another person's treasure***!"

Question? Is it time for change? **We suggest the answer is yes.**
The reason, **relevance**.

Archiving Level 0 data solves two fundamental problems:

1. Conceptually, level 0 data never "revises"
2. Can provide a plethora of customised data products for different needs
3. Standardized data QA and processing sanctioned by Scientific Community

The Big Question: Is WMO data "operational"?

Data Quality & the Data Sponsorship Statement

The “Data Sponsorship Statement” (DSS) is a fundamental descriptor of many aspects of data quality, written by the data producers (PI).

Aspects of a DSS:

- How data are measured and collected
- Network (site/platform) information and updates on any relevant changes such as obstructions etc. (where applicable)
- Compliance with SOP’s, DQO and other standards
- Calibration and revision histories
- Contact information
- DSS can act as catch-all for auxiliary information beyond data file specifications.
- Direct reference between data version and supporting DSS information.
- Version is specific and meaningful to Agency, not assigned by data centre.

Expected Outcomes

An open data initiative – **what is the new role of the data centres** – is there a new world view of data centres?

Hopefully we can establish a new mechanism for communication of ideas and future collaborations – **terms of reference**? That is, find areas of congruency and consensus on issues

If we all have to change a little for the common good, then is the change worth the effort?

In the near-term the WOUDC seeks two things:

1. From Data Producers – use of version numbers in their data and have an accompanying DSS (Improves Data Quality)
2. From Data Consumers – need citation (recognition) of the IP for each data set – viability of the data centres and collection process need this

Expected Deliverables

Perhaps a follow-up meeting in another 1.5-2 years to discuss progress

A meeting report and a list of recommendations authored by this group to be forwarded to the WMO-GAW Secretariat. As well as others? If so, who?

Any other deliverables that you wish to add?

In Closing …

What is "the agenda" for this meeting?

We want to bring together like-minded members from similar, yet diverse communities who understand our concerns and issues in order to learn from your experiences.

This perspective can often be different from the scientists who are often driving the efforts of the centres.

We are not here to "move or sell" a particular viewpoint, quite the opposite.

We are willing to share our experiences, openly and freely, but we are here to listen to your ideas and thoughts

Proposed Break-Out Session questions

1. Do you think the current system of data submission (Level 1, processed data “reports”) adequately represents the data **to the** user community? (Examples adequate temporal/spatial resolution)

2. Should we be archiving Level 0 data?

3. How can data consumers contribute more to the data archives?

4. How can we better react to the concerns of our communities? Can we share some technology or best practices? If so, what are they?