Joint Environment Canada and WMO Expert Team on World Data Centres (ET-WDC) Meeting

12-14 May 2010, Toronto Canada

Meeting Report

Executive Summary

This data centre experts meeting, held in Toronto, Canada, brought together the World Meteorological Organization / Global Atmosphere Watch (WMO/GAW) programme Expert Team on World Data Centres (ET-WDC) as well as other experts who operate or contribute to other large data inventories and archives housing data pertinent to the WMO-GAW programme. The meeting was divided into two parts, the first part focusing on general discussions and presentations among the various experts examining ways to share knowledge, experiences and further develop these systems and networks of data as the community needs evolve. The second part focused mostly on the goals and objectives of the ET-WDC group pertinent to its terms of reference and tasks in support of the GAW Strategic Plan. The meeting Recommendations, a List of Participants and the Meeting Agenda are given in Annexes I through III respectively.

Table of Contents

WELCOME – B. McArthur (Environment Canada)

Bruce McArthur, Chief, Experimental Studies Section of Environment Canada (EC) opened the meeting by welcoming everyone and began his presentation by mentioning the importance of the data centres. In the style of a library, but instead of holding books, the data centres are holders of knowledge that may not be issued today or tomorrow but many years from now. McArthur also mentioned the recent Nature article written by Jonathan Shanklin looking back 25 years ago to the discovery of the "ozone hole". Although satellite data had detected the low ozone values, it was the breakthrough paper co-authored by Shanklin using ground-based data that received world

attention. The message to be emphasised, data centres are important stewards of these ground-based data. McArthur concluded by wishing all the participants a fruitful and productive meeting.

Part 1: General Discussion on Data Centres (12-13 May 2010)

Session Chairs: E. Hare, D. Worthy and E. Carty
Rapporteur: E. Hare

Introduction – E. Hare and E. Carty (WOUDC-Environment Canada)

Ed Hare and Edward Carty from the World Ozone and UV Data Centre, Canada gave an opening presentation on the rationale and purpose for the general "experts" meeting on data centres. The rationale stemmed from a similar style meeting back in 2007 in Dübendorf, Switzerland that brought together IGACO participants which called for better collaboration and integration of data sets from the various networks and data archives. The question of whether data centres can deliver the various products and respond to requests from data "consumers" remains an issue. Hare also mentioned that the WOUDC has been serving the ozone (and more recently the UV) scientific communities for nearly 50 years, but asked the question: Is it time to set a new course? The purpose of the meeting then, was to explore the various approaches to resolving data centre issues through knowledge sharing, best practices and lessons learned from the various "extended" communities. Some of these issues relate to:

- the fact that multiple data centres do similar but not necessarily complete data archiving
- the requirements and/or needs for archivngarchiving Level 0 "raw" data versus Level 1 "data reports" versus Level 2+ custom "gridded" data
- the notion that Level 1 data are always evolving, thus, translators are time consuming and inconsistent and as a result there is inconsistent versioning and handling of revisions etc.
- the lack of a standard 'metadata' model (consistent data dictionary) and no standard 'data content' model (consistent data dictionary)
- the development of next generation of data "deployment"
- Data push vs pull,
- Data files vs Data bases,
- Binary vs ASCII etc.

The different perspectives of "Long-term" archives versus snapshot databases OR project-based archives

Two main meeting outcomes were proposed:

- Recognition and citation of Intellectual Property from data consumers (enhances viability of data producers and data centres)
- Enhanced use of Version and Data Sponsorship Statements from data producers (enhances data quality and extensibility)

Evolving influences have arisen, such as the use of the Internet, affecting the role of WMO Data Centres perhaps beyond their initial "scope" and original mandate; to serve as long-term data archives. The current WMO-GAW Strategic Plan: 2007-2015 certainly reflects the increased responsibilities.

Ed Carty then presented "Challenges to Data Managers" as a new paradigm asking the question, are we still relevant? As the GAW Data Centres were conceived many years ago, and with new technologies emerging, is it relevant to keep the "world centre" aspect intact? Handling revisions is a large task, complicated when duplicate data exist. Perhaps as a group we should consider

WEB2 type re-development of WMO Data Centres, which may involve a wholesale GAW Data Centre re-conception and re-development. How do we address the ongoing proliferation of data discovery portals, and how do we exploit the use of distributed databases?

Then Carty gave a brief presentation on the difference between Research and Operational needs. Research based data seeks discovery and innovation, and is willing to accept errors and to learn from these exceptions. In the research domain, the responsibility for data quality and production of 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 move from research to operational, these processes can potentially be offloaded to centralized 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). The main message was, in either scenario, that intellectual property of the data owners must be respected and preserved. He concluded with a brief discussion on the use and storage of Level 0, primary data.

Hare concluded the introductory presentation by briefly explaining the concept of a Data Sponsorship Statements, similar to the NDACC Metafile description, as a form of data quality reporting. Expected outcomes from the meeting included:

- 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 not worth it?

Meteorological Service of Canada's Data Management Initiative – Tony Colavecchia (Environment Canada)

Tony Colavecchia from Environment Canada began his presentation asking the question: why is data management important? Data are the foundation of EC science, products and services and data management problems are limiting achievement of results. Fixing these problems will contribute to more results, achievements of priorities and benefits for stakeholders, primarily Canadians.

A goal is to have data assets that are processed in real-time, are of known quality and effectively stored to facilitate easy discovery, access, and exchange. Environment Canada priorities include prediction renewal, warnings and alerts and dissemination/access to the Weather Office and climate data on-line and the management of and access to data assets. At EC, there are some 450 applications that are used to manage data gaps and to manage data in general. Colavecchia discussed duplicate data, inconsistencies, and the large "spider web" of data that requires coordination. Much of the data within this "web" are based on previous "products" and as a result, reprocessing down the road can sometimes lead to errors. It was suggested that reprocessing level 0 (primary) data might be the best approach to eliminate these problems.

The Environment Canada data management initiative is called the Data Management Framework (DMF) and it is working toward a leadership role in managing data. The focus used to be on quality data, now it is focussed on managing "known" quality data. The exchange of data requires standards and so the DMF is being designed to consolidate how data are managed analogous to fixing the plumbing, it is an infrastructure challenge.

Thus, dissemination and access to data fixes the "plumbing" by attaching the design to newer and better technologies. The roadmap is to continue the development of the DMF through means of renewal of data management systems which include real time, near real time, archiving, data acquisition, monitoring asset management and software retirement of legacy systems. The process begins in the short term with data sharing, by creating data discovery portals such as GeoConnections, capable of ingesting third party data including network strategies. In the longer term, it will explore options for linking databases: ontologies, data models – a semantic web. Use of native quality assessment checks will also be investigated.

Data sharing standards require accepted formats, need Metadata data sets (in XML mostly) – data discovery, access, identification of data quality and constraints – use of ISO standards including, but not limited to, ISO19115.

World Radiation Monitoring Center (WRMC) and some other activities at the Alfred-Wegener-Institut – Gert König-Langlo (Alfred-Wegener-Institute)

Gert König-Langlo from the Alfred-Wegener-Institute, Germany, gave a presentation on the World Radiation Monitoring Centre (not to be confused with the World Radiation Data Centre in St. Petersburg, Russia). These data sets are 1-minute averages collected from 47 stations that provide Global, Diffuse, Direct, and Downward Long-wave solar radiation measurements. Some of the stations also measure and monitor other radiation components such as UV, Reflex, Upward Long-wave as well as vertical soundings, aerosols and total column ozone. This complete coverage and data catalogue is required for satellite validation. Presently, there are 5743 station-months available.

The main WRMC objective is the archiving of uniform and consistent measurements throughout the Baseline Surface Radiation Network (BSRN) in order to
monitor the surface short-wave and long-wave radiative components and their changes with the best methods currently available,
provide data for the validation of satellite-based estimates of the surface radiative fluxes and
produce high quality observational data for comparison to climate models.

In 1988 the WMO proposed the establishment of the BSRN. In 1992, the BSRN began with 5 sites along side the WRMC at ETH Zürich under the direction of Prof. Atsumu Ohmura. In 2004, BSRN officially became a contributor to the Global Climate Observing System (GCOS) and by July 2008, after 15 years of nearly continuous operation at ETH Zürich, the archive moved to Alfred-Wegener-Institute (AWI) in Bremerhaven, Germany under the direction of König-Langlo.

The new WRMC-BSRN system (http://www.bsrn.awi.de) utilises an infrastructure called PANGAEA. PANGAEA is a publishing network for geo-scientific and environmental data (http://www.pangaea.de/). PANGAEA guarantees long-term availability of its content through a commitment of the operating institutions. PANGAEA follows the "Recommendations of the Commission on Professional Self Regulation in Science for safeguarding good scientific practice". The policy of data management and archiving follows the Principles and Responsibilities of ICSU World Data Centers (http://www.ngdc.noaa.gov/wdc/guide/gdsystema.html) and the OECD Principles and Guidelines for Access to Research Data from Public Funding (http://www.oecd.org/document/55/0,3343,en_2649_201185_38500791_1_1_1_1,00.html). Each dataset can be identified, shared, published and cited by using a Digital Object Identifier (DOI). PANGAEA is also used for the "Word Data Center for Marine Environmental Sciences (WDC-MARE)" (http://www.wdc-mare.org/) hosted at AWI. One can link tables, which offers easy access to any dataset using the PANGAEA system. PANGAEA presents well-defined metadata for any dataset, requires no login, and presents the data in different formats (ftp, text, html). It also utilises

Google Earth overlays and the Data Warehouse offers averaging of long time series. Many applications have been developed to assist users with graphing, file format conversion and GIS mapping capabilities.

König-Langlo suggested that the various data centres improve the interactions between archives, since there were too many places to send data files. Perhaps an XML metadata representation conforming to ISO standards is a start. He also supports the use of these processing centres and the WMO/WIS initiative.

König-Langlo then briefly reviewed work done by the AWI with specific attention given to two sites (Georg Forster and Neumayer) stations with long data time-series. Data from these stations are transmitted to several different international data networks or archives, such as: Global Telecommunication System (GTS), Baseline Surface Radiation Network (BSRN), Network for the Detection of Atmospheric Composition Change (NDACC), Global Atmospheric Watch (GAW) and the World Ozone and Ultraviolet Radiation Data Centre (WOUDC), a real-world example of the work involved in submitting data to multiple data centers.
In closing, König-Langlo offered that in terms of the interactions between archives, the PANGAEA system can be harvested from other archives using an XML metadata model that conforms to ISO standards. Finally, he proposed the WRMC become a candidate for a Data Collection and Production Centre (DCPC) within the WMO Information System (WIS).

As part of the ensuing discussion, König-Langlo pointed out that data that already have a DOI number cannot be deleted and will follow the history of the data submission. Are there costs associated with DOI? The answer is yes, the cost recovery system is being considered through user fees.

Aura Validation Data Centre (AVDC) – Christian Retscher (NASA-GSFC)

Christian Retscher from NASA-GSFC, USA gave a brief overview of the Aura (satellite) Validation Data Center (AVDC) which was established in 2004 to support the platform-wide validation activities (airborne missions, ground-based, balloons, other satellites) of the four Aura instruments (HIRDLS, MLS, OMI, TES). The AVDC also supports the Aura instrument teams, NRA PIs (Aura and other), NASA campaigns, ESA PIs, NDACC PIs, and independent validation contributors, in the validation and improvement of Aura data products.

The centre has been operational since 2005 and currently has 350+ registered users with approximately 9TB of downloads in the last year and about 500 GB of data sets along with about 6 TB of correlative satellite datasets. These data sets include: all Aura L2 data from DISC, preliminary, experimental and complimentary satellite datasets, Tropospheric ozone residual data and other L3 datasets (AIRS, Scisat ACE, NOAA 16-18 SBUV v8 profiles, Envisat GOMOS, MIPAS, SCIAMACHY, GOME2. The archive staff also maintains campaign archives (many related to Aura) such as SAUNA (1&2), WAVES, TMF NO2 campaign, ARCTAS.

Calibration and validation (Cal/Val) support includes Level 2 and Level 3 data sub-sets and collocations. Sub-setting is updated as Aura L2 data becomes available:
All OMI products (HDF5 and ASCII)

- O3: 570 sites
- Aerosol: 328 sites, including all current Aeronet sites
- NO2: 609 sites
- UV: 174 sites
- SO2: 165 sites

MLS, HIRDLS and TES
- O3, T, H2O at NDACC sites and other key profiling stations
- Subsetting of non-Aura data

MODIS, GOME2, Envisat and campaign and regional sub-setting on request

There is also direct PI support (mainly in sub-setting and data conversion), campaign support, tools and documentation that are available on-line. The centre uses the Generic Earth Observation Metadata Standard (GEOMS) and HDF formatted data files. They also provide conversion of NDACC NASA AMES data into AVDC/EVDC format. Data are converted into HDF for ease of use (reading and writing) but rely on the PI's for the data quality. There are also Aura Scientific Team and Working Group documentation and presentations available.

GEOMS allows for interoperability between data centres through enabling remote query, catalogue replication, data ordering and/or systematic mirroring. Data Centers with correlative observations use a single data format such as AVDC, EVDC (CALVAL, Envisat), NDACC and several EC Campaigns. Initiatives ongoing are to establish joint data exchange protocols and single sign on. Some standardized formats allow custom use of the data; for example, a station can request the pixel directly above the station.

Retscher then discussed the Generic Environment for Cal/Val Analysis (GECA) project which aims at delivery of
- Expanded harmonized metadata
- Study of standards supporting interoperability between validation data centers
- A validation data center implementing these standards, also interoperable standards for satellite data archives
- Open-source data conversion tools

Open-source building blocks (libraries) for collocation algorithms (both for the users local use and for the GECA server.)

The main components of the GECA Validation Data Centre (GVDC) and end-user analysis toolboxes allow Cal/Val analysts and Campaign Coordinators to coordinate cal/val activities, identify collocations and retrieve correlative data files and to analyze these correlative and satellite data files using proven and traceable cal/val analysis techniques. A Quality Information and Action Protocol (QAIP) allows investigators to identify and investigate data quality issues and to submit or query quality information

By way of clarifying the term 'single sign-on', Retscher mentioned that one only needs credentials for one data centre and this is shared between collaborative centres. AVDC has three staff, but there are other people from other centres and much is shared between these centres.

NOAA ESRL Carbon Cycle Observing Network – Ken Masarie (NOAA-ESRL, USA)

Ken Masarie from the NOAA Earth Systems Research Laboratory, USA presented the NOAA ESRL Carbon Cycle Observing Network. This network received almost 500,000 discrete measurements in 2008 alone, with data streaming in daily. Masarie likens the work of the network as including the roles of data producer, data consumer, AND data centre. There are ground-based and aircraft-based components to the network where discrete measurements are made. There are also "quasi-continuous" measurements that began in 1974 and "Tall Tower" measurements began in 1992. There are daily downloads of high-frequency data from 35 detectors.

He mentioned three “take home messages”:

- open data policy – supported by full metadata
- cooperative program – it is only there because we have people who are cooperating (this may change once there are financial costs associated with carbon emissions) – it will acknowledge the PI and the collaborators
- Metadata for all of us

Approximately 50% of the CO_2 and CH_4 contributions to the WDCGG are from the NOAA network with annual updates occurring every August for the discrete surface and quasi-continuous data and quarterly updates for the quasi-continuous tower data. Recent requests for up-to-date data include 71 requests for 2009 and 21 for 2010.

The main challenge is accommodating users with differing or competing needs. There are the needs of the PI’s in the laboratory, mostly involving logistics, but there are also research needs which encompass a wider user audience. Data use is being well documented and logged. Data are made available “on the doorstep” so data consumers are invited to come and use the data.

It is important to acknowledge the collaboration information (beyond the PI), as it is the right thing to do. Ongoing developments include full disclosure and open data, explicit acknowledgement, measurement and observation comments. Still an outstanding issue remains: how to handle “free text” information? Perhaps moving toward ISO19115 compliance?

An example of the collaboration agreement is now embedded within each data file and is given below in Figures 1 and 2 respectively.

EXAMPLE 1

```
# File Content:
#
#   Please refer to the species-specific README file in the
#   appropriate directory folder at ftp://ftp.cmdl.noaa.gov/ccg.
#
# Contact:
#
#   Compound: CO2
#   Thomas J Conway
#   tel: (303) 497-6681
#   email: Thomas.J.Conway@noaa.gov
#
# ********** COLLABORATORS **********
#
# NOAA thanks "Chinese Academy of Meteorological Sciences [CAMS]"
# without whom these measurements would not be possible.
#
# ********** RECIPROCITY AGREEMENT **********
#
# Use of these data implies an agreement to reciprocate.
```

Figure 1. Example of the collaboration agreement within each file.

EXAMPLE 2

```
#
#  ************ COLLABORATORS ***************
#
# NOAA thanks "DOE Environmental Energy Technologies Division at
# Lawrence Berkeley National Laboratory" without whom these measurements
# would not be possible.
#
# URL: http://calgem.lbl.gov
#
# The program at Walnut Grove is a collaborative effort with
# the Department of Energy's Lawrence Berkeley National Laboratory.
#
# The California Energy Commission is funding the project through
# its Public Interest Energy Research Program.
#
# WGC & STR Principal Investigator:
# Marc L. Fischer, Staff Scientist
# Atmospheric Science Dept.
#
# Lawrence Berkeley Nat. Lab.
# MS 90K-125
# 1 Cyclotron Rd.
# Berkeley, CA 94720 # MLFischer@lbl.gov email
# 510-486-5539 phone
# 510-486-5928 fax
#
# If the data are obtained for potential use in a publication
# or presentation, LBNL and NOAA should be informed at the outset
# of the nature of this work. If the LBNL-NOAA data are essential
# to the work, or if an important result or conclusion depends
# on the LBNL-NOAA data, co-authorship may be appropriate. This
# should be discussed at an early stage in the work. Manuscripts
# using the LBNL-NOAA data should be sent to LBNL and NOAA for
# review before they are submitted for publication so we can insure
# that the quality and limitations of the data are accurately
# represented.
#
#  ************ RECIPROCITY AGREEMENT ***************
```

Figure 2. Example of the collaboration agreement within each file.

Masarie then discussed ongoing developments that may impact the WDCGG, such as

- full disclosure / open data
- explicit acknowledgment of collaborators
- measurement and sampling comments

GMD working with Ted Haberman (NGDC) towards ISO 19115 compliance

The WDCGG currently provides only a subset of information that is available from the NOAA data centre. Masarie emphasised that some information relating to the data is not yet included in the WDCGG dataset, and that this additional information needs to be added. It has to be checked whether the SAG GHG will handle these issues, but it is important that a new data submission be made to add these data from NOAA to the WDCGG.

Jörg Klausen asked if the WDCGG staff asked for assistance with these recommended changes, would NOAA staff be prepared to assist them? The answer was yes. Masarie suggested as well that staff at the centres also have to react to other needs from groups like the NDACC. He acknowledged that the community as a whole has not been representing the metadata properly so the stakes are high to get this aspect right, in part to protect the PIs. What does this look like, the forging of new territory. Data may need to be reformatted or re-assigned, but always should be included, or perhaps stated differently, never be lost.

Klausen commented that the description of data quality is not well described in the ISO19115 standard (some of this may be described better in ISO19115-2), but other aspects are well described. Jacquie Witte asked how the communities and data centres were assisting the developing nations. She also commented that this aspect is an important part of NASA's outreach and the WMO capacity building efforts.

Sustaining Arctic Observing Networks (SAON) and International Polar Year (IPY) Activities – David Hik (University of Alberta, Canada)

David Hik from the University of Alberta gave a presentation on the Sustaining Arctic Observing Networks (SAON) initiative and International Polar Year (IPY) activities in Canada. The SAON initiative is "*a process to support and strengthen the development of multinational engagement for sustained and coordinated pan-Arctic observing and data sharing systems that serve societal needs, particularly related to environmental, social, economic and cultural issues*". SAON was convened by the Arctic Council following the Salekard Declaration in 2006 and renewed following the Tromsø Declaration in 2009.

Hik presented the SAON vision: *"that users should have access to free, open and high quality data that will realize pan-Arctic and global value-added services and provide societal benefits*". In order to attain this vision, SAON's goal is to enhance Arctic-wide observing activities by facilitating partnerships and synergies among existing 'building blocks', and promoting sharing and synthesis of data and information. Hik sees these activities as being both a data provider and data user.

He then asked the question: How do we manage all this information? He suggests it should be done through free and open access to data. Other questions emerge: What type of support will be required? What is essential and necessary for future investments in arctic data management? There have been five workshops held since 2007 that have produced four recommendations:

- The Arctic Council should lead efforts to ensure a sustainable pan-Arctic observing system.
- Arctic Council member states should commit to sustaining and enhancing current observing activities and data and information services.
- Arctic states are urged to increase inter-governmental cooperation in coordinating and integrating Arctic observing and data management activities.
- Arctic issues are of global common concern and open for scientific study by all states, therefore Arctic Council member states are urged to welcome non-Arctic states and international organizations as partners in sustaining and improving Arctic observing capacity, and data and information services.

SOAN may also be viewed as an IPY legacy, there are many Arctic and Global networks and platforms (both Ground-based and satellite) that in the long-term, can provide value-added services and benefits in terms of science as well as policy and decision making.

Current priorities include:

- Inventory of Observing Networks (Inventories of established networks and data archives using a standard format. Focus on long-term networks initially (see the SAON web site at http://www.arcticobserving.org/.) New and updated information will be added to this list on an ongoing basis and the list will be expanded to include other observing and data management activities).
- Facilitate data access, archiving and sharing (A forthcoming *State of Polar Data* Report reviews the current state of technology and support for discovering, accessing, and sharing polar/Arctic data. SAON and IPY Data Management Committees will host a joint workshop at Oslo IPY OSC in June 2010 to promote interoperability of observing and data management systems and identify improvements such as a "union catalog" of data sets).
- Promote Community Based Monitoring (A subgroup of SAON SG, coordinated the gathering of information on existing from local/traditional knowledge activities as well as supporting map-based registry for the SAON inventory and website. This included collaboration with the Inuit Circumpolar Council and other indigenous peoples organizations).
- Explore funding and agency cooperation (One of the most important steps toward realization of sustained Arctic observations is to obtain the views and support of the many funding and implementing organizations that deal with observations on the Arctic region).
- Recommend an institutional framework.

Hik concluded the presentation with a statement that SAON building blocks = SAON partners. It is an operational observing activity; with a defined point of contact and a process for regular dissemination of information (a Web site and/or newsletter). There is a public mechanism for obtaining information (metadata) about the observing activities and data, appropriate data quality control procedures are in place and the principle of free and open data access is being followed to the maximum extent possible. Plans are in-place for both medium and long-term data archiving at nationally- and/or internationally recognized data repositories.

Plenary Group Discussion – Data Producers (Originators) and Consumers (Users) of Data

An open plenary discussion was started with three issues to consider.
Future directions/needs/uses of data centres (what functions can/should they provide, etc)
Data quality assurance (data inclusion, uncertainty assignments, etc)
Coordination amongst data users and providers (i.e. data use policies, appropriate acknowledgements, etc)

As the open discussion ensued, further questions arose. For example, what are some of the constraints facing data centres? Beyond the obvious financial and personnel requirements there are political issues. As an example, Europe promotes project based studies and requires that data go to a specified data centre, but it is likely a "regional" centre established solely for that project. Klausen asked whether data centres "get" data (i.e "pulled" by the centre staff) or receive them (i.e. "pushed" by the data producers)? Carty responded that people are often quite willing to share data, often because of political pressure.

Klausen suggested the use of the Guide to to the Expression of Uncertainty in Measurement (GUM; cf. http://gaw.empa.ch/glossary.html) as a beginning for a dialogue on consistent data quality control vocabulary and assessment. Margins of uncertainty can vary, but uncertainty can always be defined. Masarie asked, How certain is your scale? with reference to analytical

uncertainty and atmospheric variability. Suda suggested that this issue be discussed in relevant scientific communities, so that the data centres include this information into the archive strategy. Klausen responded that perhaps each WMO SAG should discuss these issues and make recommendations on how to include this at a centre. Masarie warned the groups that uncertainty still represents different things to different communities.

Witte mentioned, as a data end user (consumer), the gaps in data reporting, anomalous data and outliers remained one of her biggest problems. Colavecchia offered that "local knowledge" is required to engage the producers to best describe the data and any "outliers", analogous with the "logbook" approach. Klausen maintained that, as a data consumer he did not really want to challenge / have to question the data centre to find out about data quality; data quality statements must be included with the data. Colavecchia agreed that challenging data producers and making it too difficult for them may lead to slow or non-existent submission of data. König-Langlo commented that when he took over the BSRN he eliminated the former data quality flags, suggesting the use of data quality assessment tools that allow a user to gauge the quality of the data in the archive. The centre would then simply provide these tools to the end user. Markus Fiebig suggested that one either should get access to primary (level 0) data or one has to rely on the SAG to assist the data centre staff in defining the quality flags. The challenge is to get the producer of data to conform or adhere to recommendations such as following Standard Operating Procedures and meeting Data Quality Objectives.

Witte commented that in her experience, addressing these issues requires a very good working relationship with data providers and the personal contacts. Hare and Carty agreed that this is the "backbone" of the WOUDC operation. Klausen re-iterated that any information about the data must be included within each data set. Hare offered that harmonising the version number between data centres of the same "theme" perhaps would resolve the problem of multiple versions or redundancy. Van Bowersox suggested that each record have a latest update identifier and that he would like to reformat all the precipitation chemistry data to unify the data sets into one format with one identifier. Jeannette Wild commented that there was value in having data in one location, yet understands that if it resides somewhere else (for example at the WOUDC), then people can go there and get access to broader data sets that are provided in a single format. Masarie also mentioned that the NOAA data centre (and database) is dynamic and constantly changing.

The final portion of Part I of the meeting involved **breakout sessions** to review the following issues:

- 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)
- Should we be archiving Level 0 data?
- How can data consumers contribute more to the data archives?
- How can we better react to the concerns of our communities? Can we share some technology or best practices? If so, what are they?

The breakout groups were:

Group #1	**Group #2**	**Group #3**
Geir Braathen (WMO)	Ken Masarie (NOAA)	Markus Fiebig (NILU)
Doug Worthy (EC)	Jörg Klausen (Empa)	Kazuto Suda (JMA)
Van Bowersox (NOAA)	Gert Koenig-Langlo (AWI)	Jeannette Wild (NOAA)
Jacquie Witte (NASA)	Chul-un Ro (EC)	Christian Retscher (NASA)
Dan Chao (NOAA)	Senen Racki (EC)	Sangeeta Sharma (EC)
David Hik (U. Alberta)	Anne Thompson (PSU)	Edward Carty (EC)

Facilitators: T. Colavecchia and E. Hare

Each group reported back to the plenary:

Group 1 – Rapporteur, G. Braathen

Issue 1
- Submit metadata in a standardized way. ISO 19115.
- Standardized units, e.g. ozone partial pressure measured by ozonesondes
- Standards for naming of stations
- WMO standards
- GAW standards
- NDACC standards
- Assess data gaps
- Create links to other data centres, e.g. links from WOUDC to the NDACC DHF

Issue 2
Yes, where possible, but with certain conditions

Issue 3
- Give feedback on suspicious data.
- Data users should give feedback on how data is used
- List of publications that use the data from the archive
- One could establish a wiki

Issue 4
- Decadal assessments of data.
- Learn where there are data gaps.
- How to improve data sets?
- Is the station network adequate
- Data format translators and data readers
- E.g. the format translators used at AVDC

Group 2 – Rapporteur, K. Masarie

Recommendations :
- Level 0 data to be archived in a central location – long-term, but access restricted

- A need for PC data assessment, collaborate and bring data together in a central location to enable the assessment – thus, this group has laid the ground work for the PC data centre – key was "regional" data centres that contributed to this assessment – there was good communication between centers and this included strong collaboration – in effect, the community came together to bring large sets of data into play for a central purpose – the assessment
- Recommend to consider a distributed data management model – it remained open what standards are required to properly "manage" this distribution or inter-operability
- Consider use of a dedicated journal for the publishing of "data" employing an open review process – for example – Earth System Science Data; *http://earth-system-science-data.net/)* -- that require a solid description and location of the data and offer a chance for data producers to get credit for their data sets.

Group 3 – Rapporteur, E. Carty

Recommendations:

- L0 data useful at WDCs, primarily for long-term storage of the data in one place, and WDCs must demonstrate their capability of providing this service.
- L0 data archives are not for dissemination of primary data, and not for reprocessing without the permission of the data owner.
- include documentation, methods, etc. as far as possible
- Encourage data consumers to provide feedback to WDCs
- need tools to implement this; WIKI / social networking
- usefulness depends on resources (time & money) of WDCs
- Investigate OAI-PMH as framework for "on demand" metadata sharing
- No consensus on metadata; likely need some standard beyond ISO19115
- WDC's have important role as long-term persistent public or member archives

Following this open discussion and breakout groups, the plenary began to draft a set of recommendations submitted primarily to the WMO-GAW Secretariat but also addressing other individuals and institutions/agencies. The recommendations are listed in **Annex I**.

Part I of the meeting was adjourned.

Part 2: WMO-GAW ET-WDC meeting (13-14 May 2010)

Meeting Chair: J. Klausen
Rapporteur: M. Fiebig

Day 1, afternoon of 13 May, 2010

ET-WDC chair Jörg Klausen, Empa, Switzerland, opened the meeting and welcomed the participants, including a brief introduction of ET-WDC in the GAW system of bodies for the guest experts. The agenda was presented and accepted. Markus Fiebig was designated as meeting rapporteur.

The meeting continued with presentations on the state of each WDC, beginning with the presentation on WDCPC by Van Bowersox, NOAA, USA. The **WDCPC** has recently been re-established at the University of Illinois in collaboration with the US National Oceanic and Atmospheric Administration. The data centre functionality is in the process of being established. SAG PC and WDCPC are currently working on a precipitation chemistry assessment report, which will be published as a journal article. The SAG also seeks to make available supplemental maps and information and will pursue posting this information on WDCPC, GAWSIS, and other appropriate web sites..

The presentation on **WRDC** by Anatoly Tsvetkov, MGO, Russia (by means of teleconference) summarised the coverage of the data collected and highlighted the statistical quality assurance procedures applied to the data. A question was asked on how metadata are submitted to WDRC? The metadata, are requested and inserted manually.

In the course of the presentation on the **WDCGG** by Kazuto Suda, JMA, Japan, the question of how to implement the WIS Discovery, Access and Retrieval (DAR) metadata was asked, i.e. how should GAW WDCs as DCPCs of WIS exchange these metadata, and which functionalities should be included in the interoperability of WDCs and GAWSIS?

The presentation on the **WOUDC** by Ed Hare highlighted the use of “Data Sponsorship Statements”, i.e. short descriptions following the data on which QA measures have been taken. Also, the importance of following up on data consumers adhering to the data policy was stressed, and that any data used are properly cited with the data centre and data originator as the source.

Since the representative of **WDC-RSAT** could not be present due to overlapping previous commitments, the information submitted to the meeting was presentated by Ed Hare. WDC-RSAT is using Digital Object Identifiers (DOI) to make data readily quotable. Data having received a DOI has to be archived permanently. If the data are changed for whatever reason, then a new DOI must be issued

The presentation on **GAWSIS** by Jörg Klausen stressed the ambition to include AERONET, SKYNET, and EANET in the database, and mentioned the need to continuously eand improve and expand the information on data quality in GAWSIS.

Klausen then gave an introduction to the **WIGOS / WIS** process. The WMO Integrated Global Observing System (WIGOS) is meant as an umbrella for the WMO OingPobserving programmes as well as co-sponseredsponsored programmes (GCOS and others), and should establish a WMO corporate identity for them. The WMO Information System (WIS) is the infrastructure that WIGOS will use for data and metadata sharing. Two WIGOS/WIS pilot projects are currently being implemented under the auspices of the Commission for Atmospheric Sciences (CAS):

1. NRT data delivery (ozone and AOD);
2. Improving the interoperability among the WDCs and GAWSIS via implementing ISO compliant metadata representations; and a portal to satellite data and related information (implemented through WDC-RSAT).

In the discussion that ensued, it was stressed that each WDC may need WMO support to procure national funding to implement all these initiatives. Geir Braathen clarified that in the context of WIGOS, all the GAW WDCs need to care about is the data discovery aspect. It was suggested that GAWSIS should take on the task to assist WDCs in being WIS-compliant. A common user / data protocol is not necessary since not all data in WIS will be accessible to everybody, i.e. data repositories may restrict access. WIS components such as GAWSISDCPC would facilitate the discovery of the data.

The question of having a common user authentication system was discussed, but not pursued further at this point.

Klausen pointed out that the pilot project (Item 2 above) requires that each WDC provide their metadata at least to GAWSIS that would then assist them being in compliance with ISO 19115

based on a WIS metadata profile to be agreed. The WDC-RSAT already complies with this standard. The question was raised whether the implementation profile should only contain the mandatory fields or all metadata held by the WDCs. Both alternatives find supporters. It was suggested that a specific example of an ISO 19115 implementation profile be made available (Note: GAWSIS provides this already, see below). Klausen concluded with the acknowledgement that the mandatory fields are required to be ISO compliant, but more metadata is needed for data to be useful to the users, including information on data quality.

Fiebig raised the question how data and access policy is upheld for NRT data cached by the GISCs. Braathen answered that this should be part of the GISC functionality, but more information on this and other issues needs to be given by the responsible persons at WMO.

Day 2, 14 May, 2010, ET-WDC expert team meeting continued

Klausen opened the meeting with a brief introduction to ISO 19115, focussing on Table 3 of the standard, and presented the implementation profile of GAWSIS. So far, it contains only the mandatory items of the standard, which is however more than enough for data discovery. Filename conventions for the XML-files containing the metadata are not yet specified. Also, whether one metadata file per data file or per dataset should be created needs to be defined along with a filename convention. Braathen proposed to include only the mandatory metadata items in the first phase while working on an extended implementation profile. The vocabularies for species and methods used by GAWSIS are also not standardised, but the terms used could be easily translated into any standard if such were put forth. Masarie proposed to install working groups for defining the vocabulary used within the metadata. None of the existing vocabularies (CF, IUPAC, AVDC) seem to be sufficient. Mapping vocabularies used at different centres is what GAWSIS effectively does at present, but it was acknowledged that this can be difficult due to ambiguities and should be avoided if possible. It was recommended that the SAGs be involved in the process of creating uniform vocabularies.

The meeting participants then re-established the break-out groups and were tasked with discussing the following questions:

1. What is the way forward for the WDCs and GAWSIS?
2. How to include info on data quality in the XML metadata?

The groups reported back to the plenary with the following reports:

Group 1:

- Data quality assessment should be kept in the hands of the PI or agency submitting the data.
- Data centres could provide quality assessment based on compliance with SOPs, calibration history, etc.
- Compliance between WDCs, common vocabulary, and inter-operability should be worked on.
- Both, small and large size WDCs have been shown to have advantages. Constraints by the variety of funding sources, a characteristic of the voluntary WMO system, have to be kept in mind.

Group 2:

- Metadata should not only facilitate data discovery, but should also document the data for use in the future.
- Reference to publications, standard methods, SOPs used, and audits should be included.
- One option is defining a few well-defined methods for data processing for each type of measurement.
- Including calibration histories may be of advantage for reprocessing data.
- The ET-WDC should ask the SAGs to provide standardised vocabularies and metadata elements needed to use data 100 years from now. Klausen was tasked to initiate this process which was deemed to take up to 2 years to complete.
- Include GAWSIS in a distributed European GISC.

Group 3:

- Importance of the personal relationship between data centre and data providers was stressed.
- A web-forum as means for user interaction could be established.
- WDCs could cater to the younger generation by being present on Facebook and Twitter.
- The task of keeping the WDCs sustainable is a task of managing change.
- Regardless of size, WDCs need expertise on instruments, either by the centre manager or by expert groups.

The meeting continued with a discussion on how a station is included in the GAW programme, and what the status of mobile stations should be in GAW. So far, stations contributing data to a WDC have been registered as regional stations by default. Klausen presented a proposed application process for joining the GAW network that was the result of discussions with the WMO Secretariat (Doc_6.1). According to this proposal, OPAG EPAC JSC (or selected members of it) would decide on acceptance of stations. Fiebig proposed to accept any stations accepting the GAW data quality principles as contributing stations, but have an approval process for regional or global stations, which was largely accepted by the group. There was a view that the definition of regional and contributing stations should be reviewed with a view to clarifying how the NMHSs be involved in the classification of such stations. It was briefly discussed whether old, inactive stations should be removed from GAW. In view of being able to discover existing data from now inactive stations, this was not deemed to be reasonable. There were also views that such labels should be applied carefully based only on data submission to WDCs, to avoid losing the motivation to data reporting. Also, it was discussed whether acceptance in the GAW programme should be based on data quality, what the review process should look like, and whether a new category for mobile stations is needed. It was agreed that stations entering GAWSIS by way of submitting data to a WDC or other data centre should be registered as ‘Contributing’ by default. ET-WDC continued the discussion on the status of mobile stations through e-mail exchange after the meeting was adjourned. From this discussion, the recommendation to drop the implicit attribute of ‘fixed’ for the three existing categories of ‘Global’, ‘Regional’, and ‘Contributing’ emerged. ‘Mobile’ stations will be listed in GAWSIS as a fourth category, as has been done by a number of data centres. It was generally agreed that mobile platforms need three-letter codes like GAW IDs for fixed stations, but there were different views on whether such codes should be assigned to platforms or programmes, and how such cases should be accepted in the GAW programme or GAWSIS.

Agenda item 7 (Update of tasks for the GAW Strategic Plan) had to be deferred to later exchange among the members of ET-WDC because of time constraints.

No other business was brought to the attention of the meeting.

The chair closed the meeting and thanked the host for the excellent organisation.

ANNEX I – Recommendations

- Recommend that each WMO SAG provide guidance to data producers on how to describe and quantify uncertainty within their data. Visit the following URL: http://gaw.empa.ch/glossary.html and reference the GUM document.
- Recommend that the various data centres continue to collaborate with one another with better defined terms of reference.
- Recommend that the identity of the primary source of the data always be preserved and communicated.
- Recommend that Level 0 (primary) data be archived for long-term storage. Access to these data would be restricted and granted only under specific conditions. Also, that the WMO World Data Centres consider performing this service, but that it be contingent on demonstrating a capability to perform this service.
- Recommend that data centres post reports of assessments and inter-comparison studies to assist data consumers in assessing data quality.
- Encourage the development of a metadata profile within the ISO19115 standard satisfying the needs of the relevant atmospheric communities.
- Recommend that the SAGs be part of the process in defining the data content vocabularies.
- Recommend the documentation of the QA procedures/processes used at the WDCs for acceptance of data.

ANNEX II – Participants List

Participant	Affiliation	Data Network
ET-WDC Managers		
Jörg Klausen	EMPA	GAWSIS, Chair ET-WDC
Geir Braathen	WMO	WMO
Anatoly Tsvetkov	MGO	WRDC (**by teleconference**)
Kazuto Suda	JMA	WDCGG
Van Bowersox	NOAA/U. Illinois	WDCPC
Ed Hare	EC	WOUDC
Markus Fiebig	NILU	WDCA
Guest Experts		
Edward Carty	EC Consultant	WOUDC
Jeannette Wild	NOAA	NDACC
Gert Koenig-Langlo	AWI	BSRN
Jacquie Witte	NASA-	SHADOZ
Christian Retscher	NASA	AVDC
Ken Masarie	NOAA	NOAA CCG
Dan Chao	NOAA	NOAA CCG
Doug Worthy	EC	WMO Global Station Alert + SAG GHG
Sangeeta Sharma	EC	Aerosols
Senen Racki	EC	WMO Global Station Alert
Tony Colavecchia	EC	Data Management for Env. Canada
David Hik	U. Alberta	IPY
Chul-un Ro	EC	Prec. Chem.

ANNEX III - Agenda

Wed. 12 May 2010

General Data Centre Managers meeting

Time	Topic	Presenter
0900-0915	Welcome by the Chief of Experimental Studies Section (ARQX)	B. McArthur, Chief (ARQX)
0915-0930	Intro by the WOUDC on the rationale for the meeting, outline the proceedings and expected outcomes from the first meeting (May12-13)	E. Hare & E. Carty
0930-1030	Brief presentations from guest experts	Session Chair: E. Hare
0930-0940	Data Management Framework at Environment Canada	T. Colavecchia
0940-0950	BSRN and Activities at AWI	G. Koenig-Langlo
0950-1000	GHG Activities at NOAA	K. Masarie
1000-1010	AVDC	C. Retscher
1010-1020	Sustaining Arctic Observing Networks (IPY)	D. Hik
1030-1100	**Coffee break**	
1100-1215	Group discussion on problem areas in data reception, coordination with data authors & providers and issues of data QA	Session Chair: D. Worthy
1215-1330	**LUNCH**	
1330-1500	Group discussion on problem areas in data delivery to clients, coordination of efforts (areas of congruency) among the various data centres and third party service delivery (satellite validation, modelling communities)	Session Chair: E. Carty
1500-1530	**Coffee break**	
1530-1700	Break out sessions (3 groups) to discuss areas of improvement to service delivery, assistance to data providers and potential future collaborations	Facilitator: E. Hare
1700	**Adjourn for the day**	
1800-	Ed will accompany the group downtown (for those who wish to see the city centre) – location of dinner TBD	

Thurs. 13 May 2010

0900-0915	Introduction for the day's activity	E. Hare and J. Klausen
0915-1030	Break out sessions (3 groups) continue to discuss next steps	Facilitator: T. Colavecchia
1030-1100	**Coffee break**	
1100-1130	Break out groups report to the plenary	Session Chair: E. Hare
1130-1230	Group to draft a set of recommendations and 1st meeting adjournment	Session Chair: E. Hare
1230-1400	**LUNCH**	
	Start of WMO ET-WDC Managers meeting	
1400-1415	1. Approval of the Agenda, denoting the meeting Rapporteur	Session Chair: J. Klausen
1415-1530	2. Presentations from the ET-WDC Data Managers	
1530-1600	**Coffee break**	
1600-1730	3. Review current status of ET-WDC WIGOS/WIS Pilot Project implementation and agree on future activities	
1900-2130	Group dinner hosted by Environment Canada	

Fri. 14 May 2010

0900-1030	4. Metadata items and vocabularies maintained at each WDC and GAWSIS: observed variables, sampling, inlets, methods of analysis 5. Metadata concerning data quality (traceability, instrument history, etc.)	Session Chair: J. Klausen
1030-1100	**Coffee break**	
1100-1200	6. Registration of 'new' stations in GAW, clarification of station types and status, representation at each WDC and GAWSIS	
1200-1330	**LUNCH**	
1330-1500	7. Review and update tasks of relevance to ET-WDC and WDCs as listed in the GAW Strategic Plan2 8. Any other business	
1500	**Meeting adjournment**	