4th International Digital Curation Conference
December 2008

Data Curation in Climate and Weather:

Transforming our ability to improve predictions through global knowledge sharing

Clifford A. Jacobs,

Steven J. Worley,

Jacobs/ National Science Foundation, USA

Worley/ National Center for Atmospheric Research, USA

July, 2008

Abstract

Climate and Weather are of increasing interest to the scientific community and the general public. Data curation and stewardship are essential building blocks in the science community's quest to better understand how natural climate and weather systems behave and how activities of human civilization are altering the natural system. Rudimentary observations of the atmosphere and ocean have been collected for over one hundred years and proxity measurements of the climate can trace our planet's climatic history for millions of years. These observations coupled with the rapid advances in technology, such as powerful computers, rapid access to massive amounts of data, and satellite observations, have allowed innovative techniques to be used to understand and predict the planet's climate and weather.

The 4th International Digital Curation Conference taking place in Edinburgh, Scotland over 1-3 December 2008 will address the theme *Radical Sharing: Transforming Science?*.

Introduction

Something is not right with the weather and climate. It seems to be changing. But how do we know for sure? Can we remember what the local weather was 12 days ago? Our colleagues in the UK might have an advantage by going with probabilities: overcast, cool and raining. We have known for years that the climate of the Earth has undergone considerable variations over eons and that the climate and weather systems are very "noisy." Natural variability is a dominant mode and this mode can mask long-term trends. We have weather data going back hundreds of years and proxy climate data going back millions of years. But, observations of the atmosphere are very uneven in space and time. For example, there are very few *in situ* observations over the vast regions of the southern ocean and only recently have satellite measurements made it possible to adequately capture the atmospheric and oceanic state in this and other remote regions on the earth.

To try to get a better understanding of atmospheric behavior at regular time and space intervals, scientists have developed value added data using a method known as reanalysis. Past observational data are used with advanced assimilation techniques and forecast models to produce retrospective data sets typically on a global grid at approximately 1^0x1^0 spatial resolution and at 6 hour intervals. The earliest reanalysis starting date is 1948 and frequently other reanalyses begin in 1979, which coincides with the advent of good satellite data coverage.

In addition to reanalyses which support climate change assessment and retrospective weather studies, there is a worldwide activity underway to capture and curate the output from global ensemble weather forecast models. This effort is focused on research to improve weather forecasts for the benefit of all nations. In this case, real-time data streams are continuously collected from 10 different national numerical weather prediction centers, systematically organized and provided to the research community in convenient forms.

The curation of these data is essential for understanding natural and anthropogenic climate change and improving our forecasting capabilities. This paper provides a brief overview of data archived at a leading research center in the US. In addition, some of the new approaches to value-added data curation in weather and climate are presented in sections that following. The paper concludes with a discussion of sustainability of data curation.

Trends within the US on Data and Data Preservation

National Science Foundation (NSF) has long supported the preservation and curation of research data through awards to institutions and consortiums of universities. Prominent examples are the protein data bank[1] which dates back to the early 1970s and climate and weather data maintained at the National Center for Atmospheric Research[2] (NCAR) which has received support since the mid-1960s. However, these examples are the exceptions. NSF's strategic goals[3] —Discovery,

[1] Protein Data Bank website: *http://www.rcsb.org/pdb/home/home.do*
[2] NCAR website: *http://www.ncar.ucar.edu/*
[3] National Science Foundation Investing in America's Future Strategic Plan FY 2006-2011: *http://www.nsf.gov/pubs/2006/nsf0648/nsf0648.jsp*

Learning, Research Infrastructure — are predominately achieved through support to the university community in the form of grants and cooperative agreements lasting one to five years in length. Often these awards produce digital data of value beyond interests of the original investigator(s). Continued support to maintain, curate, and make available these data offers a challenge to the investigator in NSF's highly competitive peer review award environment.

The National Science Board (NSB), (NSF's governing body) recognized the importance of digital data produced under NSF sponsorship and held a series of workshops starting in 2003. These workshops culminated in an NSB report[4] which states "..[NSB] recognizes the growing importance of these digital data collections for research and education, their potential for broadening participation in research at all levels, the ever increasing .. NSF investment in creating and maintaining the collections, and the rapid multiplication of collections with a potential for decades of curation." Following this report, NSF sponsored a workshop to further examine the issue of long-term stewardship of Digital Data sets. In a often cited report[5], the workshop lays out its objectives to explore "issues concerning the need for **new partnerships** and collaborations among domain scientists, librarians, and data scientists to better manage digital data collections; necessary **infrastructure development** to support digital data; and the need for **sustainable economic models** to support long-term stewardship of scientific and engineering digital data for the nation's cyberinfrastructure."[6]

The notion of cyberinfrastructure was thrust into prominence within the US with the NSF report from the Blue-Ribbon Advisory Panel on Cyberinfrastructure[7](January, 2003)[8]. This report set the stage for NSF to issue a strategic plan for cyberinfrastructure for the 21st century[9] (March, 2007). One of the four thrusts in the strategy focuses on Data, Data Analysis, and Visualizations which "sets forth a framework in which NSF will work with its partners in science and engineering … to address data acquisition, access, usage, stewardship and management challenges in a comprehensive way." Early implementations of the strategy are in the form of two solicitations: Community-based Data Interoperability Networks (INTEROP) (NSF 07-565 Solicitation)[10] (July 2008) and Sustainable Digital Data Preservation and Access Network Partners (DataNet) (NSF 07-601 solicitation)[11] (January 2008).

The preservation of digital data is a US government-wide issue that has interest across government agencies. The Interagency Working Group on Digital Data (IWGDD) constituted under the auspices to the Committee on Science of the National

[4]Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century: *http://www.nsf.gov/pubs/2005/nsb0540/*
[5] To Stand the Test of Time: *http://www.arl.org/bm~doc/digdatarpt.pdf/*
[6] Bolding from original report
[7] Revolutionizing Science and Engineering Through Cyberinfrastructure *http://www.nsf.gov/cise/sci/reports/atkins.pdf*
[8] The Cyberinfrastructure concept was well established in Europe in the late 1990's and influenced the development of the U.S. incarnation of the concept
[9] Cyberinfrastructure Vision for 21st Century Discovery *http://www.nsf.gov/pubs/2007/nsf0728/index.jsp*
[10] *http://www.nsf.gov/pubs/2007/nsf07565/nsf07565.pdf*
[11] *http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.pdf*

Science and Technology Council[12] was formed recently to address the various aspects of digital data. A report is in preparation by the IWGDD that will provide a strategy to ensure that digital scientific data can be reliably preserved for maximum use in catalyzing progress in science and society. The discussion of the digital data issue characterizes the problem in five dimensions — three of space, one of time, and one of digital information and communication technologies. It is clear that the digital data, its use, and preservation are becoming an increase important part of the US investment portfolio.

Many of characteristics outlined in the cited reports above --- data preservation, curation, infrastructure, partnerships, and sustainability --- were addressed decades ago by the Division of Atmospheric Sciences at NSF by the establishment and long-term support of NCAR. At NCAR there is a well established computational and data infrastructure which is available to the science community. This infrastructure and set of services was developed over the years andare in response to community needs.

NCAR Data Archives

The Mass Storage System (MSS) is tape based and the primary data archive facility at NCAR. In early 2008 the MSS surpassed 5 petabytes of data storage (including second copies), with a net growth rate of 80 - 100 terabytes per month. The data content reflects the diverse NCAR research portfolio and more than 40 years' history. Some of the data it stores originate from field experiments and observations: international climate records from the past 100 years include data from weather stations, ships, planes, and satellites. However, a majority of the data from global climate-simulation models, mesoscale weather models, and other earth-science models are run on the NCAR supercomputers. The MSS is exemplary of a curation management success. Since its instantiation in the middle 1980s the MSS has evolved through multiple tape media changes, scaled to accommodate ever more productive supercomputers, and has successfully been adapted to new ways scientists do their research. The MSS has a remarkable data preservation record.

The NCAR Research Data Archive (RDA) is a comparatively small (currently 246 TB, less than 5% of the MSS total size), but very important, part of the MSS stored data. The RDA has been curated by the staff in the Computational and Information Systems Laboratory for over 40 years, and as such contains reference datasets used by large numbers of scientists. The RDA contents are long-term atmospheric (surface and upper air) and oceanographic observations, grid analyses of observational datasets, operational weather prediction model output, reanalyses, satellite derived datasets, and ancillary datasets, such as topography/bathymetry, vegetation, and land use. The RDA is not a static collection; it is now over 580 datasets with about 100 routinely updated and 10-20 new ones added each year.

The RDA relevance to good curation practices can be viewed from two different perspectives: user data services and archive content development. It is interesting to note that from the RDA prespective, data services and archive content development are inextricably linked in a positive feedback loop aimed at improving in both activities. The notion of data services and the feedback from the user community does not appear to be addressed directly in the The Digital Curation Centre Curation Lifecycle

[12] Website for NSTC: *http://www.ostp.gov/cs/nstc*

Model[13].

In FY07 about 5400 unique users were provided 100 TB of data through various primary access pathways, the NCAR MSS, public servers on the web, one-time special requests prepared for individuals, and the TIGGE[14] archive (Figure 1a-b).

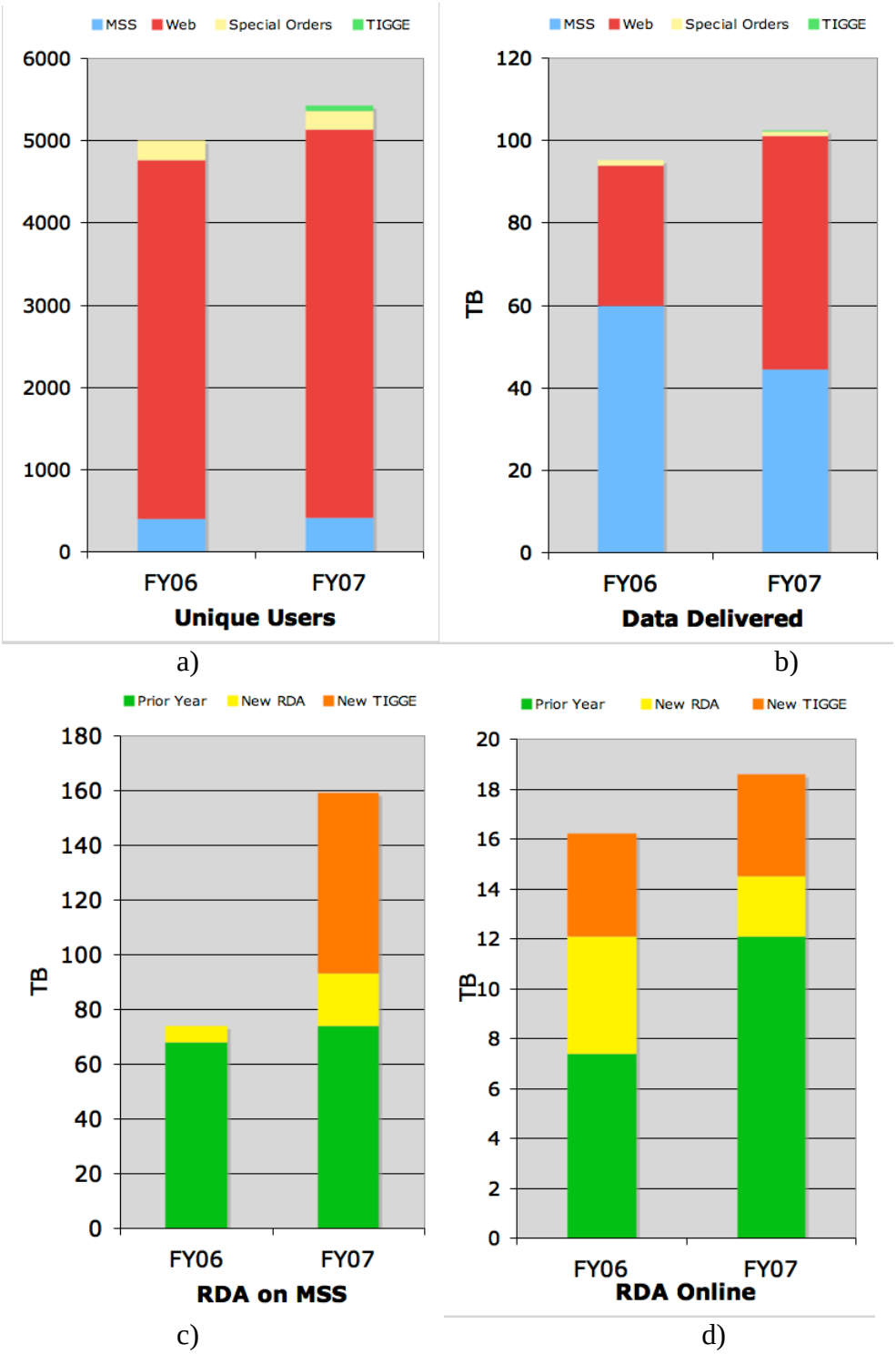

Figure 1
Service and growth metrics for the Research Data Archive (RDA) during FY06 and FY07, a) number of

[13] Data curation lifecycle: http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf
[14] See TIGGE section below.

unique users separated by modes of access, b) the amount of data provided separated by access pathway, c) amount of data in the archive on the MSS showing the annual growth, and d) amount of data on the public servers (a high-demand subset of the RDA archive) showing the annual growth.

The largest user group is associated with the web access pathway (4700) and almost equal amounts of data are delivered from the MSS and over the web, 55 and 46 TB, respectively. This indicates the value of the RDA curation activities to the worldwide research community.

A simple measure of data content development is archive growth. The RDA expanded by nearly a factor of two in FY07, from 74 to 159 TB (Figure 1c). TIGGE is part of the RDA, but is shown and discussed separately[15] because it alone added 66 TB and it represents an important curation activity in and of itself. Although this factor somewhat overshadows the 19 TB growth in remaining part of the RDA this is also large when contrasted to the 6 TB growth in FY06 (Figure 1c). The most demanded datasets from the RDA are available online through publicly available web servers. The FY07 online data is 19 TB (Figure 1d)

A copy of RDA is maintained off site at the San Diego Supercomputer Center (SDSC) to ensure the presevation of this valuable resource. Efforts underway to develop a system to keep the SDSC archieve current as the NCAR archive is updated.

Reanalyses

Atmospheric reanalyses are a main feature within the RDA and were intended to be, and have become, a very valuable data resource for a wide variety of climate and weather studies. By combining many types of atmospheric observations with advanced data assimilation and forecast models a “best possible” 3D estimate of the atmospheric state over extended time periods is achieved.

Reanalyses are supported by many historical data sources that have been curated over time. As an illustration the major sources of atmospheric profile data include wind only soundings beginning in 1920 (Figure 2). These are augmented with soundings of temperature, humidity, and wind beginning in 1948.

[15] See TIGGE section below.

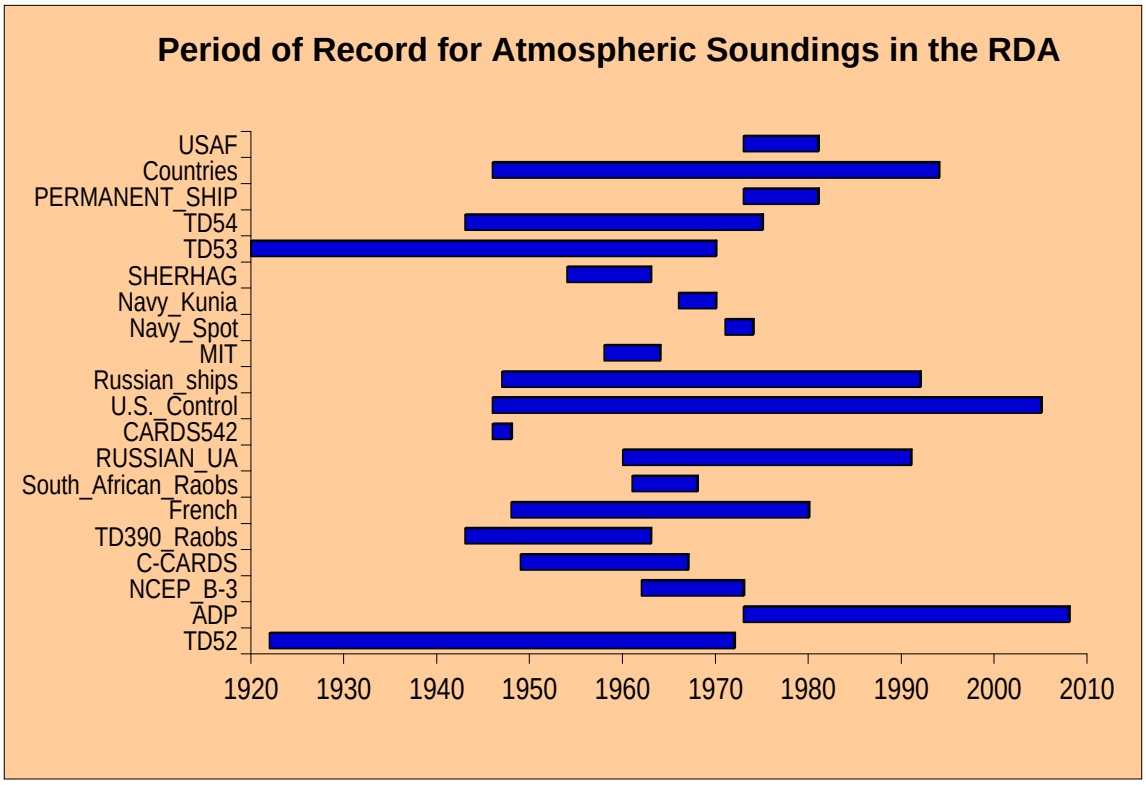

Figure 2 Period of record for major sources of atmospheric sounding data in the NCAR Research Data Archive. The name of each dataset source is shown along with date range covered.

Each model and data assimilation scheme used is somewhat different[16], but a critical aspect is that the code remains fixed and therefore any atmospheric parameter changes observed in the output **cannot** be attributed to the model. The first models used for reanalysis were based on the best available numerical weather prediction systems available at major weather forecast centers, e.g., U.S. National Centers for Environmental Prediction (NCEP). Since then the models and assimilations methods have improved; the computational grids are at higher spatial resolutions, more accurate model components have been added (land-surface, ocean-surface and coupling, radiation physics, etc.), different assimilations schemes have been developed and evaluated, and greater numbers of data products are created for a more diverse set users.

The data assimilated are a major consideration in any reanalysis. Significant effort is required to collect, organize, quality check, and prepare observational data for reanalysis. The observations include conventional surface and upper air observations taken at land stations and from ocean observing systems, and many satellite data streams that estimate the vertical atmospheric structure and earth surface parameters. It is important to note that the observational data used in each successive reanalysis are largely the same, but also continuously improving. Erroneous data can be detected and fixed, new sources of observed data are discovered and prepared, observing system biases are systematically corrected, and more accurate forms of satellite data are ingested as the models become more capable (e.g., using satellite radiance estimates versus vertical soundings calculated from the radiances). Assimilation of these diverse

[16] For example, the reanalysis done in the US and Europe use different models and data assimilation schemes

data types is a complex task and changes in time of these observing systems remains as a significant challenge. For example, consider the density of conventional observations over the globe in the early periods and then the merger of these data with satellite estimates beginning in the 1970s. Five major reanalyses are currently available in the RDA (Table 1).

Table 1, The most recent reanalyses available in the Research Data Archive, listed in order of production.

Name	*Time Period*		*Highest Resolution*		
	Start	*End*	*Temporal*	*Spatial Horizontal*	*Spatial Vertical*
NCEP/NCAR Global Atmospheric Reanalysis	*1948*	*2008 (ongoing)*	*6 hours*	*T62 (209km)*	*17 Plvl*
NCEP/DOE Global Atmospheric Reanalysis	*1979*	*2007 (ongoing)*	*6 hours*	*T62*	*17 Plvl*
ECMWF Re-Analysis 40-year (ERA-40)	*1957*	*2002*	*6 hours*	*T159 (125km)*	*23 Plvl*
North America Regional Reanalysis	*1979*	*2008 (ongoing)*	*3 hours*	*32 km*	*29 Plvl*
Japanese Reanalysis	*1979*	*2008 (ongoing)*	*6 hours*	*T106 (1.125 deg)*	*23 Plvl*

Many data products are available with each of these. A few notable features include:

- the earliest starting date is 1948 and three begin in 1979 which is coincident with availability of good satellite data,
- four are ongoing which means the computations have been continued and the time series is extended – this is an important asset for climate research,
- four of the reanalyses are global and one is one regional to North America,
- the temporal resolution is 6 hours (3 hours in the regional case), and
- spatially there is considerable variability, ranging from about 2.5°x2.5° (T62) to 32 km, and vertical resolution general at 17 pressure levels or more.

The value of reanalyses is further confirmed by the fact that typically there are future reanalyses planned or in the computational phase. Each new effort results in improvements in outcomes and often the output is at higher resolution in all dimensions.

TIGGE

TIGGE, THe Observing System Research and Predictability EXperiment (THORPEX) Interactive Grand Global Ensemble was established by the World Meteorlogical Organization (WMO) World Weather Research Programme with a goal to accelerate improvements in 1-day to 14-day high-impact weather forecasts. It is highlighted here because it is a major modern challenge for sustainable curation. To enable the research aimed at improving weather forecasting, three archive centers have been formed to receive, archive, and distribute ensemble model weather forecast data from 10 international Data Providers (Figure 3). NCAR is one of the TIGGE Archive Centers,

the others are the China Meteorological Administration (CMA), and the European Center for Medium-Range Weather Forecasts (ECMWF).

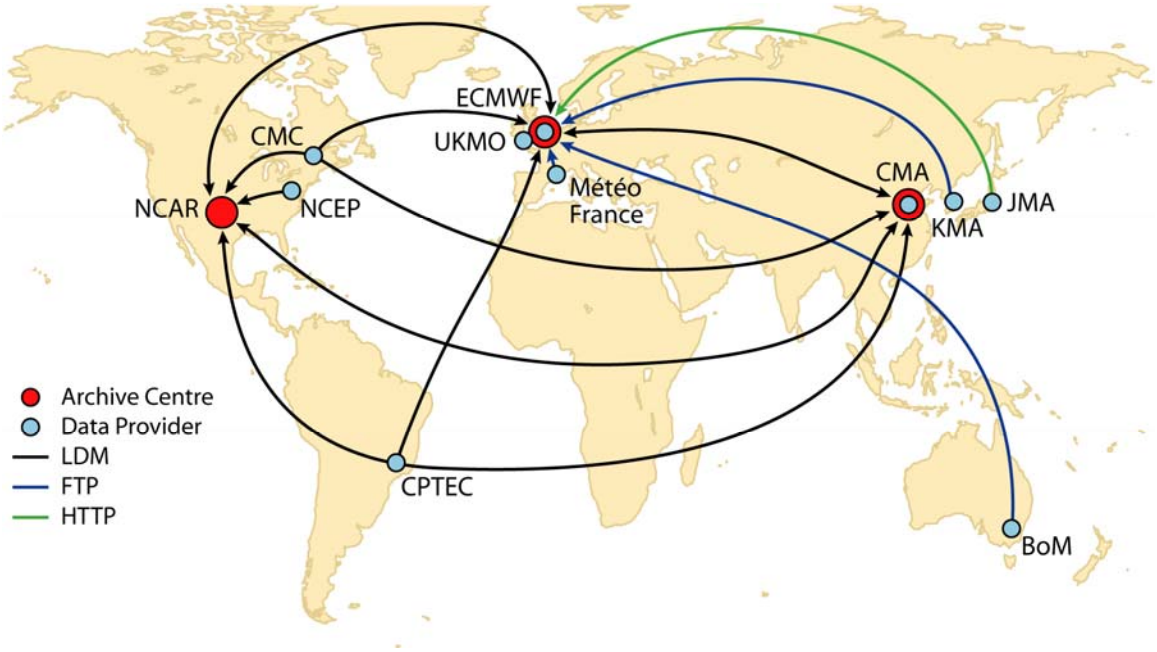

Figure 3 Schematic of TIGGE data flow and transport methods used between Archive and Data Provider Centers. European Centre for Medium-Range Weather Forecasts (ECMWF), United Kingdom Meteorological Office (UKMO), Japan Meteorological Agency (JMA), National Centers for Environmental Prediction (NCEP), China Meteorological Administration (CMA), Centre Meteorologique Canadien (CMC), Bureau of Meteorology Research Centre (BoM), Meteo France (MetF), Korea Meteorological Administration (KMA), Centro de Previsao de Tempo e Estudos Climaticos (CPTEC).

The distribution of TIGGE data employee technology developed by the Unidata program[17] under NSF sponsorship. One of Unidata's functions is to develop and maintain software tools for the distribution of real-time meteorological data for research and education. One of the technologies of these technologies is the Internet Data Distribution and Local Data Manager (IDD/LDM) system. This system is the primary data transport mechanism used between the Archive Centers. However, data providers use a combination of IDD/LDM, FTP, and HTTP to deliver data to the Archives in near real-time (Figure 3). The data transport is about 1.6 M fields and 240 GB per day to NCAR. This data volume and real-time aspect force a number of curation issues. Systems must be able to recover from network outages, production delays at the providers, and unexpected and planned power down events throughout the array of data centers and data providers. Without recovery, the archives would have missing data and therefore would be an incomplete record which would severely impact potential research.

Providing data access to such a large and dynamic archive is also a challenge. The NCAR TIGGE portal[18] provides access to the most recent two weeks of data with selection multiple options. Interfaces are available to browse the archive and select either forecast files for download, or individual parameters, spatial subsets from the

[17] The Unidata Program Centers website: *http://www.unidata.ucar.edu/*
[18] NCAR TIGGE portal: *http://tigge.ucar.edu/*

global domain, and user specified grid resolution across multiple centers. Users have the option of selecting either GRIB2 or netCDF data format. Users with NCAR computing privileges have access to the full period of record. For others, the offline files can be restaged for online access.

Sustainable data curation

There is a spectrum of critical elements necessary for sustainable data curation. Implicitly, successful curation must include stewardship so the archived data can be accurately maintained over long periods and always remain accessible and understandable to the users. There are several main elements necessary to sustain data curation:

- **Robust data storage facilities** (hardware and software) that are capable of accurately handling data migration across generations of media. Over the past two decades the tape media in NCAR's mass storage system has been upgraded about every three or four years. The total volume of the system, order 6 Petabytes, and the need for it to constantly serve the users dictates that data migration to new media must not interfere with daily operations. Under these size and operational constraints the time needed for migration almost matches the technology cycle for new media. In other words, data migration is a continuous aspect of operations.
- **Backup plans**, that are tested, so irreplaceable data are not at risk. Unintended data loss can occur for many reasons: some major causes are: poor stewardship leading to the loss of metadata to understand where the data is located and documentation to understand the content, physical facililty and equipment failure (fire, flood, irrecoverable hardware crashes), accidental data overwrite or deletion. All irreplaceable data should be stored in at least two physically seperate locations and managed with distinct systems that are not susceptible to the same faults.
- **Science-educated staff** with knowledge to match the data discipline is important for checking data integrity, choosing archive organization, creating adequate metadata, consulting with users, and designing access systems that meet user expectations. Staff responsible for stewardship and curation must understand the digital data content and potential scientific uses. Archives supported this way offer the greatest benefit to science, because they are logically organized and easy to use.
- **Non-proprietary data formats** that will ensure data access capability for many decades and will help avoid data losses resulting from software incompatibilities. These formats must be fully documented to the byte level. It is certain that computing operating systems and application software will continue to improve and change. The format choosen for the data should not be tightly couple, or dependent, on either of these. By using a few standard data formats over long periods of time and placing the burden of access on the evolving applications we can preserve the capabiltiy to understand the data that are stored.
- **Consistent staffing levels** and people dedicated to best practices in archiving, access, and stewardship, e.g. as documented in the OAIS Archive Reference Model[19]. No matter how well an archive is documented, a great deal of

[19] Reference Model for an Open Archival Information System (OAIS):

information is held by individuals who have performed the stewardship and curation work. The value of this human knowledge-base cannot be underestimated. Over time there will always be more questions asked about the data, problems will need to be evaluated and corrected, and past experiences often make new problems easy to solve. People with 10 or more years of stewardship practice are very important assests for any curation system.

- **National and International partnerships** and interactions greatly aids in shared achievements for broad scale user benefits, e.g. reanalyses, TIGGE. No one facility can do it all, yet the global science community needs it all. Partnerships and worldwide sharing of data and unrestricted open data access systems provide science research opportunities that are greater than any one center can provide.
- **Stable funding** not focused on specific projects, but data management in general. New projects are undoubtably important and receive much attention early on, but the best research data archives are built and maintained over many decades. There needs to be a clear understanding that sustained funding is necessary to keep a curated collection viable.

Successful curation requires a stable commitment from people, facilities, and an institutional organization. For many years the Computational and Information Systems Laboratory (CISL) at NCAR has fulfilled this role. CISL serves the computing, research, and data management needs of atmospheric and related sciences. The curation resources can be compared to the current total CISL annual budget. The complete MSS facility infrastructure and the cost for five system engineers that keep it operational account for about 11% of the CISL annual budget - again, note this serves all of NCAR and the external science community. The RDA is maintained by eight engineers at a cost of 6% of the CISL budget. These rough estimates do not account for broadly distributed network services, an array of web servers and associated administration, and many other support functions offered as part of NCAR. Taking the historical perspective with 40 years development and the staffing allocation over that period we see that over 200 person-years has been invested in RDA curation. It is impossible to estimate the great value of these data holdings.

Final Thougths

The nexus of observations, computer models and technology have proven to be an emergent paradigm in the century long quest to understand the atmosphere of Earth. This nexus has lead to major scientific advances and forms a foundational building block in deciphering the complexities of our climate and weather system. For example, the Intergovermental Panel on Climate Change (IPCC) used curated climatologically data available through the reanalyzes of the 20th century to checked the veracity of numerous computer models used to predict the climate of the 21st century by comparing the simulations made by the model of the climate of the 20th century climate to the reanalysis data.

Curation and stewardship have significantly different meanings although the terms are used somewhat interchangebly in this paper. Museums curate data by collecting and preserving the item or data it has acquired. The DCC Curation Lifecycle Model

http://public.ccsds.org/publications/archive/650x0b1.pdf

provides an excellent graphical representation of the curation process. Stewardship, on the other hand, includes all the activities that take place to make the information accessible and useable – things like reorganization, meta data harvesting and development, fixes, integrity checks etc. Stewardship is essentially a user driven activity. Curation and stewardship must be strongly bonded to ensure user productivity which is far more likely to lead to sustained support.

Within the US, researchers and research administrators are begin to gain a greater appreciation for the importance of the relationship between the modes of support, e.g., short term grants, group activities, national center, etc., and the capabilities of the awardees/stakeholders to sustain and make widely available data collected or generated during the period of research support. This issue is under active consideration by NSF and other parts of the US government.