DATA MANAGEMENT IN NOAA

William M. Callicott

July 25, 1991

ABSTRACT

NOAA has 11 terabytes of digital data stored on 240,000 computer tapes. There are an additional 100 terabytes (TB) of geostationary satellite data stored in digital form on specially configured SONY U-Matic video tapes at the University of Wisconsin. There are over 90,000,000 non-digital form records in manuscript, film, printed and chart form which are not easily accessible. The three NOAA Data Centers service 6,000 requests per year and publish 5,000 bulletins which are distributed to 40,000 subscribers. Seventeen CD-ROMs have been produced. Thirty thousand computer tapes containing polar satellite data are being copied to 12 inch WORM optical disks for research applications. The present annual data accumulation rate of 10 TB will grow to 30 TB in 1994 and to 100 TB by the year 2000. The present storage and distribution technologies with their attendant support systems will be overwhelmed by these increases if not improved. Increased user sophistication coupled with more precise measurement technologies will demand better quality control mechanisms, especially for those data maintained in an indefinite archive. There is optimism that the future will offer improved media technologies to accommodate the volumes of data. With the advanced technologies, storage and performance monitoring tools will be pivotal to the successful long-term management of data and information.

TABLE OF CONTENTS

1.0. Data Management in NOAA

Management of environmental data and information resources is becoming an increasingly visible and important issue for the scientific community. This is of particular importance to the National Oceanic and Atmospheric Administration (NOAA), which routinely measures and collects large amounts of environmental data and information in its own work, and is also officially charged with maintaining environmental records for the Nation. Through its activities over time, NOAA has become the steward of a treasury of Earth systems data and information--the most comprehensive, long-term, and up-to-date environmental description of the Earth that exists today. This treasury contains answers to urgent environmental questions facing the Nation. The success of all NOAA's scientific work, and the national priorities which it supports, depends on the accountability and accessibility of environmental data and information. These data and information must be accurate, complete, stable and fully-integrated across the spectrum of NOAA's organizational functions, and they must be made easily accessible, in a timely and cost-efficient way. Throughout this effort to provide resources to meet the NOAA missions of managing data for global change research and for enhancing warning and forecast services, there will be a continual inherent process to migrate and protect data held in the NOAA archives.

2.0. NOAA Data Management Operations:

The NOAA Centers respond to requests from a broad community of research, legal, engineering, individuals, insurance, business, consultants, and manufacturing. About 6,000 requests per year are received for digital data. Within the next two years as global change research increases, this should grow to 9,000. The current data files stored on computer readable media have a volume of 11 terabytes stored primarily on 240,000 computer tapes. There are 100 TB of GOES data at the University of Wisconsin copied in digital form on specially recorded SONY U-matic video tapes. There are analog, i.e., non-digital, holdings having a volume equivalent to over 50 TB in a digital domain. With the conversion of some of the analog data to digital format, and with new sources of digital data, the digital holdings will grow by 1996 to about 35 TB not counting source level data from GOES. The rapidly expanding quantity of data will force a different approach to managing data and information within the centers. The use of an integrated mass storage systems with appropriate file management will be essential to manage the archive, storage hierarchy and data migration processes. New mass storage hardware and software technologies will have to be developed and adopted and the current archives copied. If a mass storage system is not developed, the data centers will require immense tape storage areas and reduce the access mechanism from data granules to physical volumes of data.

In accordance with the NASA/NOAA Memorandum of Understanding for remotely sensed earth observations, data processing, distribution, archiving and related science support, EOS data is intended to be archived at the NOAA data centers. The EOS

data from prototype operational instruments used for NOAA operational purposes, will be an inherent part of NOAA's data archives. As part of the EOS pathfinder activity, NOAA is migrating the environmental satellite data from the operational polar orbiting satellites from computer tapes to 12 inch SONY optical write once, read many (WORM) disks. Also, there are selected special sensor microwave data from the Air Force Defense Meteorological Satellite Program (DMSP) of interest to EOS scientists. Initially, 8 terabytes of data will be migrated to optical disk.

The computing capacity at the centers is provided by mainframes, workstations, and personal computers (Pcs). In the aggregate, the systems do not have the capacity to handle the anticipated archive, the growth in data ingest and dissemination, and expanded quality control, analysis and reprocessing requirements in the coming years. The computational capacity will need to grow from 10 MFLOPS at the beginning of 1991 to over 300 MFLOPS by 1996, i.e., a factor of 30:1. On-line disk storage should grow from 60 GB to over 500 GB during this period. The configurations are evolving from a central processor surrounded by dedicated terminals to a fully distributed client-server architecture which can expand in response to workload demands.

2.0. NOAA Data Centers: There are three National Data Centers and over a dozen centers of data in NOAA. The data centers are structured as formal archive centers and serve at this Nation's world data centers for their respective disciplines. The following provides a description of the centers and their activities.

2.1. The National Climate Data Center (NCDC) in Asheville, North Carolina was established in 1950. The National Archives and Records Service, in compliance with the Federal Records Act of 1950, specified that NCDC's climatological records be permanently retained. It has been designated as a World Data Center (WDC)-A for Meteorology. It also operates the Satellite Data Service Division which manages the high volume satellite data. The Center is responsible for ingest, quality control, archiving, managing, providing user access, and performing analysis of data which describes the global climate system. The NCDC also supports major new programs such as the National Weather Service Modernization, Climate and Global Change, the Coast Watch Initiative, and the Level-0 Earth Observing Systems (EOS) path finder effort. In 1992, new data sources from foreign satellites will be introduced. The center has a staff of 290 full time employees (FTEs), 100 contract and 190 federal employees. Each working day results in 160 orders from 360 user contacts. Annually, about 5,000 bulletins are prepared and supplied to 40,000 subscribers. Much of NCDC's data (by physical volume) is in the form of paper records such as ship logs, and although manually accessible, is not readily usable, and they appear to be deteriorating. There are fifty thousand cubic feet of paper records at NCDC. There are also film and other non-machine readable information stored in the National Archives.

NOAA has a contract with the University of Wisconsin to collect and archive data from the NOAA geostationary operational satellites (GOES). The GOES data collected from

1978 to the present, is recorded and stored at the University of Wisconsin's Space Science and Engineering Center. To date, the GOES data are stored on 25,000 Sony U-matic beta video (19mm commercial video standard) video cassettes in 4 GB increments. Access to the data is not highly efficient because some of the cartridges are offsite in the state records office, and because the data must be reingested as a satellite readout. The GOES archive represents the largest amount of data anywhere in the NOAA system to be rescued and be made more readily accessible. To improve access, the center is engaged in a pathfinder study on mechanisms to improve the access to the data.

2.2. The National Geophysical Data Center (NGDC) in Boulder, Colorado was established in 1972. Its mission is to manage solid earth and marine geophysical data as well as ionospheric, solar and other space environmental data; and to provide facilities for World Data Center-A for Geophysics which encompasses Solid Earth Geophysics, Solar-Terrestrial Physics, Marine Geology and Geophysics, and Glaciology. The center has a staff of 60 FTEs.

NGDC has over 300 databases including 54 million (M) Ionograms, 2.5M magnetograms, 12 million flight miles of aeromagnetic data and 1O million miles of ship track data. There are 25,000 magnetic tapes partly at NGDC in Boulder and partly in Asheville at NCDC. About 2000 tapes per year arrive from originators outside of NGDC. Their goal is to keep all of the ingest tapes and maintain two additional copies for use in normal center activities (3 copies total). About 14,000 tapes have no backup. There is a requirement to copy 12,000 tapes a year for routine migration. NGDC relies on the error checking provided by the tape copying system and supplements this with printouts of beginning/end of record data and record counts.

NGDC began the NOAA CD-ROM program four years ago. Its first CD-ROM title was "Geophysics of North America". The center now has a total of 12 CD-ROMs completed or underway. This effort should change the distribution system for data from its tape orientation and probably deflect some use of the network to obtain data. During 1990, distribution of data by CD-ROMs has increased both the number of requests for digital data and the annual amount of data distributed by the center. The distribution of data by CD-ROM puts the data into a form highly convenient and useful to Pcs and workstations.

The National Snow and Ice Data Center (NSIDC) under contract to NGDC operates the World Data Center-A for Glaciology. The role of NSIDC is to acquire, archive and disseminate data relating to all forms of snow and ice. It provides data to about 500 requesters per year from a digital archive data base of about 15 GB (300 standard tapes); 7GB are from the NIMBUS-7 Scanning Multi-channel Microwave Radiometer, and 3 GB are Special Sensor Microwave/Image (SSM/I) data. Many of the NSIDC datasets are redundantly held at other NOAA data centers. A daily volume of 1 GB of data from the Defense Meteorological Satellite Program (DMSP) Operational Linescan

System (managed by NGDC) will be forwarded from the DMSP readout site to NSDIC on EXABYTE tape. A similar means is being developed to distribute a weekly data volume of 800MB between the Joint Ice Center (JIC), in Suitland and the National Snow and Ice Data Center (NSIDC) in Boulder. The NSIDC has been designated as an EOSDIS Distributed Active Archive Center (DAAC). As such, it will build up its computational and archival ability to meet EOSDIS defined mission goals. As a DAAC, the NSIDC will relocate to the University of Colorado campus.

2.3. The National Oceanographic Data Center (NODC) has been in operation since 1961 as an interagency, facility under the U.S. Navy, and became a part of NOAA in 1970. Its mission is to manage oceanographic data. It has operated as a World Data Center-A (WDC-A) for Oceanography since 1962. NODC has a staff of 85 FTEs. NODC's files include data collected by U.S. federal agencies; state and local government agencies; universities and research institutions; foreign government agencies and institutions; and private industry. Currently, NODC maintains a digital archive of both in-situ and satellite-sensed ocean data in excess of 30 GB. A potential equivalent of 10 GB of digital data are currently maintained in analog form such as data reports, manuscripts, and analog instrument recordings. With the establishment of NODC data management responsibilities for ocean observing satellites, including non-NOAA geostationary and orbiting platforms; new global collection efforts, including the Global Ocean Flux Study (GOFS), the World Ocean Climate Experiment (WOCE), and the Climate and Global Change Project; and new U. S. coastal ocean studies, including CoastWatch and the Coastal Ocean Program, the archive is expected to increase twenty fold between FY90 and FY95.

NODC has a large amount of analog data. A tablet digitizer is used to annually process 10,000 expendable bathythermograph traces (XBT) to 2 MB of data. NODC has a contract with the Navy to annually process 100,000 similar traces. Mechanical bathythermographs (MBT) on glass slides (300,000) await conversion and are expected to result in 1.5 GB of data. Acoustic Doppler Profiling, done from University and NOAA ships, is expected to be a new source of data. There are perhaps 20 ships that may be equipped with these profilers. At the present, only a few are capturing the data for archiving purposes. Each profiler should provide 10 MB per ship month. In the future, 100 ship months per year of these data could be archived.

The NOAA Coastwatch Data Management, Archive, and Access System (NCAAS) now under development will result in expansion of the archive, related quality control, and retrieval and distribution activities based on SONY 12 inch WORM Optical Disks. NODC also is responsible for the NOAA Library and there is interest in digitizing some of its data and metadata holdings.

NODC is responsible for the NOAA Earth System Data Directory, and interfacing it with the larger NASA based master directory effort. This is part of the catalog interoperability effort which is underway among other government agencies and

foreign data centers. The master directory is also needed to fit with the EOSDIS version 0 effort where the catalog interoperability will perform relevant IMS (Information Management System) functions. NOAA's master directory is VAX based, with specially written software, and the ORACLE Data Base Management System.

A prototype database system has been developed to provide NODC users with direct access to an on-line, interactive data archive. It maintains a data base of over 23 million marine observations from 310,000 ocean stations. Access to the data is obtainable through spatial or temporal searches with arbitrary combinations of instruments, platforms, and parameters. By FY 1993, NODC plans to add all of its vertical profile data (Nansen, Bathythermograph, C/STD, etc.) to the POSEIDON system.

2.4. NOAA Centers of Data include those data collection and operations elements performing observations and monitoring services as part of NOAA's recurring mission responsibilities. Listed below are the principal centers:

Discipline	Title	Location
Bathymetry, Nautical charts, Geodesy	Charting and Geodetic Services	Rockville, Maryland
Climate	Climate Analysis Center	Camp Springs, Maryland
Fisheries	National Marine Fisheries Service	Seattle, WA; Woods Hole, MA; Miami,FL; Bay St. Louis,MS; San Diego, CA
Ice	Navy Joint Ice Center	Suitland, Maryland
Lake Data	Great Lakes Environmental Research Lab	Ann Arbor, MI
Oceanography	Center for Ocean Analysis and Prediction	Monterey, CA
Oceanography	Ocean Products Center	Suitland, Maryland
Pacific Ocean Data	Equatorial Pacific Information Collection	Seattle, WA

Particle Deposition Data	Air Resources Lab	Silver Spring, MD
Sea Level	University of Hawaii	Honolulu, HI
Snow and Ice	National Snow and Ice Data Center	Boulder, CO
Tides	National Tide and and Water Level Data Base	Rockville, MD
Trace Gases	Global Monitoring for Climate Change	Boulder, CO

3.0. Digital Data Request History and Projection

The support for global change by the NSF has increased about 35% per year since 1987 and is expected to increase for FY 1992. NOAA's global change funding has roughly doubled each year since 1989. Other agencies are also increasing their global change funding. The overall funding for all agencies has increased seven times from FY89 to 1991. From this, one could expect a large increase in the number of data requests at the centers. However, in the aggregate, there has only been a modest increase in the number of requests for each of the last two years ('89 and '90), and the projections are, therefore, based on this modest rate of increase. Another view is that the secondary distribution of data from scientist to scientist may be on the increase because of the ease of transmission over networks, coupled with a desire to obtain a dataset that has had use in a familiar science project. Possibly this secondary distribution masks the size of the actual data dissemination.

The biggest impact on the number of requests and volume of data distributed has been from the introduction of CD-ROMs. This indicates that the increased use of CD-ROMs for data distribution provides a means for rapid deployment of the data among members of the research community. Secondary distribution of data from scientist to scientist may be on the increase because of the ease of transmission over networks, coupled with a desire to obtain a dataset that has had use in a familiar science project.

4.0. Mass Storage/File Management Requirements

As the amount of data increases and the NOAA mission focuses on improving accessibility of data for global change research, there is an urgent requirement to develop mass storage systems with file management software at the centers to

improve archive management, provide vastly improved access mechanisms, and to reduce the amount of media and associated space requirements. Moreover, the mass storage system is the heart of a file management system. For the immediate purpose at hand, a mass storage system should include the following:

4.1. Hierarchical File Director

A hierarchical file directory is needed that permits, as a minimum, the acceptance of UNIX file names. The directory needs to maintain the access and update history information for the file. The directory should allow for handling mixed media within a single search, for cross indexing between devices, and for recording data set utilization records for future knowledge based system applications. This directory must interoperate with a number of different data base systems passing query information through during interoperable data searches.

4.2. A hierarchy of storage levels

The mass storage system should support a hierarchy of storage levels with increasing physical capacity and decreasing performance at the bottom, and decreasing physical capacity and high transfer rates at the top (as viewed from the user client processes). At the top, this permits the evolution to direct access electronic storage, so that the mass storage becomes a truly integrated part of a computing environment.

4.3. File Management Software

The file management software should offer options for data compression. It should permit the use of checksums as an overall error control mechanism. Data conversion software should be available to migrate the data from one physical media to another, as generic files, without disturbing the data content. The software would sample files on a statistical basis reading them to verify that they were still intact and that the media had not deteriorated beyond the point where only soft data checks were obtained. In the event that sufficient degradation was detected during this sampling process, the files would be migrated to new media with a corresponding directory update. Migration would also be triggered during normal accesses whenever too many soft errors occur.

Migration of files from archival or working media to a buffer storage area would occur following the initial access to permit data to be more rapidly retrieved from the faster devices in the storage hierarchy. The migration and actual location of the data should be presented to the user/application programs in a transparent manner including, as an option, presentation to client processes in a manner simulating direct retrieval from the ingest media if desired. To accomplish this transparency, the file management software should provide for the ingest of data from existing media and distributed to: standard half-inch magnetic tapes in all densities and formats, EXABYTE, DAT, CD-

ROMs, optical disks, video cassette recording technologies, digital optical media, etc. An encapsulation of the ingested media's data should record the presence and location of permanent data errors, physical record lengths in bytes, the presence and location of ingest media specific flags, such as tape marks, end of tape flags, etc., so that upon access, the data can be handed to processing programs that need to be aware of the different media. The file management software should be able to handle any of the existing data formats and to invoke conversion routines to a standard if one is adopted. The ability to move files from the mass storage to a requester's media in the original format should be provided.

4.4. Network Access and Networking

The NOAA centers should be coupled to the internet, at internet backbone rates, and eventually to the National Research and Education Network (NREN). With 740 universities, laboratories and industrial sites now on the network, and 75 more expected in the 2nd half of 1991, the internet is widely available to the scientists involved in global change and EOS. There are 16 NSFnet backbone sites. Two of these, the University of Maryland and NCAR, are in close proximity to NOAA data centers. Where large data volume data transfers are required, conventional conveyance services would probably suffice with economic considerations determining the mode of conveyance.

5.0. Considerations

The system life under the NOAA mandate to manage data for long-term global change research purposes is open ended. The value of data increases with age for use in performing long term environmental change research. Global patterns are known to be subtle over time, even when viewed in a rapidly changing environment. Today's collection of environmental data is pitifully small and of too short a duration compared to the amount required to filter out the statistical noise over an extended time domain.

The operational requirements are influenced by incremental science requirements established as the knowledge of the relationships between instrument responses and conversion to physical units became better known from the results of research and development of more sophisticated processing algorithms. Because the development of sufficient knowledge to fully understand the earth observation responses is an accretive and repetitive process, the entire data set will require repeated reprocessing to adequately describe the data for long-term documentation and preservation.

Aggregation of similar but different and sometimes disparate data types is also an important feature to include. A well understood aggregation principle will allow for compartmenting the data across the media domain in a "most" convenient form for vastly improving the access mechanisms. This will become increasingly important as

the longevity of the archive increases. Aggregation implies some degree of redundancy, but in a positive sense, in that redundancy of particularly critical data sets reduces the risk of data loss over time.

Volume management may require compression mechanisms to reduce the over-volume of data as it ages. Critical data and information will require the application of lossless data compression where data sets are reduced in volume. When permitted, other means can be used to reduce data volume with controlled data loss as achieved through sampling, or through small-loss data compression techniques or a combination of the techniques to yield much higher compression ratios. This may particularly attractive for managing very high volume image data sampled in the visible spectra.

The technological gallop of the last several years continues to accelerate and new storage and processing technologies are obviating the need to consider destruction of cumbersome data through full scene sampling, scan sampling or reduction to gross descriptive parameters which describe the sum of the parts in abbreviated form. In order to take advantage of improved and less costly storage technologies, there is a plan to migrate the data periodically as the volumes dictate and the technology allows and with each migration to yet developed capabilities, it becomes even more feasible to consider placing all of the data in a near-line access environment. The migration process will require content processing to re-develop the cross reference inventory information to include additional content description information as part of the inventory metadata file to increasingly document the data as it ages. Data migration is an essential element in developing a never-ending data life for the sake of offering an extended time baseline data set essential for detecting global scale changes.

A wide variety of media will be used for distributing data and information to users. The large number of formats used by the NOAA data centers means that many conversion procedures will be needed. It would be better not to convert the data in the archives themselves, but to write procedures which can be invoked in a demand fashion. In this way, the data can be left in its original form while confidence is gained in the accuracy of the conversion. If any problems arise in the converted data, the original data will not be contaminated. The problem becomes one of reworking the conversion routine and alerting previous users of the defect rather than trying to fix a partially scrambled dataset. The downside is that there will be an additional processing cost when the data is requested. Another problem with the standard data format concept is that many researchers who submit data to the data centers will probably never conform to a standard format. Insistence on a format may become an impediment to releasing the data to a center for distribution.

6.0. Assumptions and Constraints

Factoring today's technology advancement timescale for the purpose of being both realistic and conservative, the period of migration to exploit new technologies and avoid system obsolescence to extend the validity of the data and information is established at no less than every 10 years. A suggestion was recently made that the migration rate be a function of the expected half life of the medium used. For the 3480 tapes, the manufacturers agree that 10 years of full performance life should be expected, thus the half life for migration purposes would be five years. The criteria for accepting a new technology as a candidate for data migration is; improved archival qualities, the per data byte storage cost must be one half the previous, the physical storage requirements be at least five times less, and the data transfer rate to move the data from the media be no slower than that of the older media. And, finally, the data migration step will include data processing to derive or extend content description data to be used for the purpose of validating the preservation state of the data and for reinventoring the data to add content information developed through the accretion of user knowledge and experience.

Another assumption is that all of the data will be reprocessed three times during a 25 year cycle. This reprocessing cycle may coincide with a data migration step since all of the migration will include a content review during the passage of data from one medium to another. The development of reprocessing algorithms will not be charged as a data management system task but will require that the data management system be able to put significant quantities of data on-line or at-hand for "live" ingest mode processing.

7.0. Conclusion

To continue managing data as we do today would eventually require a facility to accommodate an enormous number of media units to hold the data volumes projected for the future. As the new data continues, the added function of migrating data from degenerating media (from a systems as well as physical viewpoint), will compound the storage requirements as the annual data volume accumulates by the hundreds of trillions of data samples each year. If acceptable mass store systems are not continually developed to match the data growth and data management requirements, the logistics would be overwhelmed and the system would fall apart never to be recovered again because of the enormous cost to recover an inevitable backlog.

The data only has value to the research community if it is conveniently and efficiently accessible. If the data were placed in a warehouse environment, which would ultimately have to happen if nothing was done, it would soon lose its value and possibly its identity because of the cost to acquire it and eventually would be lost because of the huge cost to locate, ship back, copy and return the data copies as the data volume grows beyond manageable proportions. This is aside to the issue of data

loss due to media deterioration. The only acceptable solution would be through development of a system capability to provide highly efficient and sophisticated data management capabilities which would accommodate online data sources. In order for this to happen, advanced media technologies have to be employed along with advanced sophisticated data management software to eliminate the manual interfaces where possible to provide the data in a ready mode for user interaction at the subsetting level. The data value increases dramatically when placed in this type of environment as the access to the system can provide instant gratification and encourages repeated and expanded data query activities. This, in turn, accelerates the research progress and enhances the research results thus broadening the value of the data to the advancement of science and knowledge.

To physically compress the data through the implementation of high capacity media coupled with the systems capability to control and index these data in an online or near-line environment offers a significant reduction in the requirements to house the archives both in terms of physical space and recurring energy and labor expenses. The closer on line the data are placed, the less labor is required to physically mount data either in the appropriate archive slot or onto the processing system. As electronic access become more fully integrated into the system, direct labor service categories will be eliminated. A major cost avoidance to be reckoned with is the cost of adding increasing large physical facilities as the data volume grows at the projected rates. The pace of implementation of new technologies should allow shrinking of the space requirements to match the increase of data accumulating in the facilities. An indirect benefit of space compression through improved storage technologies would be realized form compressing the facilities requirements sufficiently to consider replicating the data in distributed locations as a risk reduction measure.

The broadest benefits are in terms of what the value of the data is to the world of environmental change. Without a responsibly managed data record of scientific measurements over time, there would be no baseline to objectively determine if the environment we live in is really changing, how much, and at what rate. Without these data, economies would be based on subjective opinions and in some cases, hear say. Public policy would more often than not be misguided and consequences of enormous proportions could occur to our physical well being either through economic collapse or through direct physical changes. Even today, global change scenarios portend potential devastating effects to our coastal cities and this country's agro-economies. But, do we build dikes and seek alternate water sources if we are not really sure what, if any, impacts there are? Without the data, no one knows, so any investment in mitigating a potential problem is an economic risk. The other question is; even if we know, can we afford to take avoidance action? Or better yet, is the cause due to environmental causes or due to a much broader cyclic processes. One thing is certain, there is a great potential for change based on the knowledge at hand today, and sound economic planning based on knowledge may be sufficient to avoid economic collapse. In a Nation with a trillion dollar economy, the risks are too great

not to invest an insignificantly small percent of this economy to acquire the maximum amount of knowledge possible and establish this knowledge base as soon as possible. In the case of environmental data, data is knowledge; there can never be enough data, and the data record can never be long enough. But where there is data, it must be systematically managed to be of any value at all.

DATA MANAGEMENT

WILLIAM M. CALLICOTT

OFFICE OF SYSTEMS DEVELOPMENT
DATA MANAGEMENT SYSTEMS DIVISION
NATIONAL ENVIRONMENTAL SATELLITE,
DATA AND INFORMATION SERVICE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

NSSDC CONFERENCE ON MASS STORAGE SYSTEMS AND TECHNOLOGIES
FOR SPACE AND EARTH SCIENCE APPLICATIONS

NASA/GODDARD SPACE FLIGHT CENTER
GREENBELT, MARYLAND
JULY 25, 1991

DATA MANAGEMENT PROCESSES

- INGEST
- QUALITY CONTROL
- CATALOG
- ACCESS
- PRESERVATION

INGEST

- Three discipline centers and U.S. World Data Centers:

  National Climate Data Center - Asheville, North Carolina
  National Oceanographic Data Center - Washington, D.C.
  National Geophysical Data Center - Boulder, Colorado

- Sixteen centers of data:

  Satellite Data Processing and Distribution - Suitland, MD
  National Weather Service - Silver Spring, MD
  National Meteorological Center - Camp Springs, MD
  Charting and Geodetic Services - Rockville, MD
  Climate Analysis Center - Camp Springs, MD
  National Marine Fisheries Service - Seattle, WA; Woods Hole, MA; Miami, FL; Bay St Louis, MS; San Diego, CA
  Navy Joint Ice Center - Suitland, MD
  Great Lakes Environmental Research Lab - Ann Arbor, MI
  Center for Ocean Analysis and Prediction - Monterey, CA
  Ocean Products Center - Suitland, MD
  Equatorial Pacific Information Collection - Seattle, WA
  Air Resources Laboratory - Silver Spring, MD
  University of Hawaii - Honolulu, HI
  National Snow and Ice Data Center - Boulder, CO
  National Tide and Water Level Data Base - Rockville, MD
  Global Monitoring for Global Change - Boulder, CO

NOAA CENTERS

- DIGITAL HOLDINGS INCLUDE: 11 TB ON 240,000 COMPUTER TAPES
- IN-SITU DATA INCLUDED IN ABOVE: 2 TB
- GOES DATA HELD AT THE UNIVERSITY OF WISCONSIN: 100 TB ON 25,000 SONY U-MATIC BETA TAPES
- SERVICE OVER 7,000 REQUESTS FOR DATA AND INFORMATION PER YEAR
- ANNUALLY PRODUCE 5,000 BULLETINS TO 40,000 SUBSCRIBERS
- OVER 90,000,000 PAGES OF NON-DIGITAL DATA AND INFORMATION

IMMEDIATE DIGITAL VOLUME GROWTH
(VOLUMES IN BILLIONS OF BYTES)

	1991	1992	1993	1994	1995	1996
CLIMATE DATA CENTER	520	570	840	1,300	2,130	3,420
OCEANOGRAPHIC DATA CENTER	30	160	200	260	280	370
GEOPHYSICAL DATA CENTER	450	540	650	780	930	1,110
SATELLITE DATA SERVICES	10,400	14,200	18,000	21,800	25,600	29,400
GOES DATA ARCHIVE	107,000	113,000	120,000	134,600	149,200	164,000
ACCUMULATIVE TOTAL:	118,400	128,470	139,690	158,740	178,140	198,300

ALL ENVIRONMENTAL SATELLITE SOURCES
Includes level-1 (L0+10%) plus levels 2-4 (40% of L0)

VOLUMES IN TRILLIONS OF BYTES PER YEAR

SATVOLS 6/25/91 Disk #8

SATELLITE SYSTEM	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Conventinal OBS	2.00	1.00	1.00	2.00	3.00	4.00	5.00	5.00	5.00	5.00	6.00	6.00	6.00	6.00	6.00	8.00	8.00	8.00	8.00	8.00	8.00	10.00	10.00	10.00
NOAA-D (5/91)	0.55	1.10	1.10	0.55																				
ERS-1 (9/91) (ESA)	0.03	0.11	0.11	0.11																				
NOAA-I (12/91)		1.10	1.10	1.10																				
JERS-1 (3/92) (NASDA)		0.08	0.16	0.16	0.08																			
TOPEX (6/92) (NASA)		0.08	0.16	0.16	0.16	0.08																		
GOES-I (10/92)		0.93	9.27	9.27	9.27	9.27	9.27	9.27																
GOES-J (6/93)			4.64	9.27	9.27	9.27	9.27	9.27	4.64															
NOAA-J (12/93)				1.10	1.10	1.10																		
NOAA-K (7/94)				0.48	1.20	1.20	0.72																	
RADARSAT (7/94) (CAN)				0.05	0.16	0.16	0.16																	
ADEOS (3/95) (NASDA)					0.08	0.16	0.16	0.16																
NOAA-L (7/96)					0.48	1.20	1.20	0.72																
NOAA-M (2/97)							0.90	1.20	1.20	0.30														
TRIMM (3/97) (NASA)							0.05	0.11	0.11	0.11														
GOES-K (10/97)							1.85	9.27	9.27	9.27	9.27	9.27	4.64											
GOES-L (6/98)								4.64	9.27	9.27	9.27	9.27	9.27	4.64										
EPOP-A (9/98) (ESA)								5.06	20.26	20.26	20.26	15.19												
EOS-A (12/98) (NASA)									58.20	58.20	58.20	58.20												
JPOP (12/98) (NASDA)									0.18	0.18	0.18	0.18												
NOAA-N (9/99)									0.30	1.20	1.20	1.20												
NOAA-O (4/01)											0.60	1.20	1.20	0.60										
EOS-B (6/01) (NASA)											29.10	58.20	58.20	58.20	29.10									
EPOP-B (9/02) (ESA)												4.05	20.26	20.26	20.26	16.21								
GOES-M (10/02)												1.85	9.27	9.27	9.27	9.27	9.27	7.42						
EOS-C (12/02) (NASA)													58.20	58.20	58.20	58.20	58.20							
GOES-NEXT (6/03)													5.37	10.73	10.73	10.73	10.73	10.73	5.37					
JPOP-B (12/03) (NASDA)														0.18	0.18	0.18	0.18							
NOAA-P (4/04)														10.13	20.26	20.26	20.26							
EPOP-C (9/06) (ESA)																5.06	20.26	20.26	20.26	10.13				
EOS-D (9/06) (NASA)																14.55	58.20	58.20	58.20	58.20	29.10			
NOAA-Q (4/07)																	10.13	20.26	20.26	20.26				
GOES-NEXT-P (10/07)																	2.15	10.73	10.73	10.73	10.73	10.73		
EOS-E (12/07) (NASA)																		58.20	58.20	58.20	58.20	58.20		
GOES-NEXT-PP (6/08)																		5.37	10.73	10.73	10.73	10.73	10.73	
NOAA-R (4/10)																				10.13	20.26	20.26	20.26	
EPOP-D (9/10) (ESA)																				14.56	58.20	58.20	58.20	58.20
EOS-F (9/11) (NASA)																					14.55	58.20	58.20	58.20
NOAA-S (4/13)																							10.13	20.26
EPOP-E (9/14) (ESA)																								14.55
TERA BYTES / YEAR		2.57	4.39	17.53	24.25	24.82	26.46	28.61	44.71	108.43	103.80	134.09	164.63	172.41	178.21	154.00	142.46	197.38	199.16	191.75	200.93	209.77	226.32	167.52	161.21

- TIROS N:	1.10	Tera Bytes/Year
- NOAA-KLM:	1.20	Tera Bytes/Year
- GOES Composite:	1.97	Tera Bytes/Year
- GOES-I/M GVAR:	7.30	Tera Bytes/Year
- GOES-NEXT GVAR	8.76	Tera Bytes/Year
- NOAAPOP:	20.26	Tera Bytes/Year
- EPOP:	20.26	Tera Bytes/Year
- EOS-A Platforms:	58.20	Tera Bytes/Year
- EOS-B Platforms:	58.20	Tera Bytes/Year
- NOAA EOS Reqts	11.50	Tera Bytes/Year
- NOAA JPOP Reqt's:	11.83	Tera Bytes/Year
- Eos SAR Platform	118.26	Tera Bytes/Year
- Space Sta Freedom	0.18	Tera Bytes/Year

Accumulated in 20 years: 2.3 Peta Bytes
From 1991 to 2011

Daily Totals in MBytes for Level 1 plus Levels 2 - 4 Data (1.5 x L0):			GBytes/Day Archive Vol
- Composite Mapped GOES:	3600		5.400
- GOES-I/M bulk GVAR:	20000		20.000
- GOES Next bulk GVAR:	24000		24.000
- TIROS-N:	2000		3.000
- NOAA KLM:	2200		3.300
- NOAA Platform:	37000	(3.5 Mb/s)	55.500
- European Platform:	37000	(3.5 Mb/s)	55.500
- EOS-A Research Platform:	106304	(9.84 Mb/s)	159.456
- EOS-B Research Platform:	106304	(9.84 Mb/s)	159.456
- NOAA Reqt's off EOS:	21000	(2 Mb/s rate)	31.500
- JPOP	21600	(2.0 Mb/s)	32.400
- SAR (Special SAR S/C)	216000	(20 Mb/s rate)	324.000
- Space Station Freedom	324	(.03 Mb/s)	0.486

PRESERVATION

- NOAA MISSION TO MANAGE THE ARCHIVES ON A PERMANENT BASIS AS A NATIONAL TRUST
- ALTERNATIVES FOR PRESERVING DATA ON AN INDEFINITE BASES:
  - FIND A MEDIA THAT IS INDELIBLE INDEFINITELY
    - ... Manage access system indefinitely
    - ... Ensure media longevity (entomb media and site)
    - ... Recurring Quality control
    - ... Lifetime system maintenance
    - ... Migrate on demand
    - ... Operate archive center(s) as deep archive
  - ASSUME A 10 YEAR SYSTEM AND TECHNOLOGY CYCLE
    - ... Recurring migration of 10 year old media continually looking at developing technology advantages
    - ... Keep two copies, one entire data set entombed, one in active archives at distributed discipline archive centers
    - ... Maintain portions of the data at the active centers on-line, the rest near-line
    - ... The master copy kept near-lie with sufficient on-line capability to service access requests and for migration processing

ARCHIVE INTEGRITY

- MEDIA PERFORMANCE CONTINUALLY MONITORED THROUGH ERROR DETECTION AND CORRECTION INFORMATION PASS BACK
  - Processing on-demand
  - Scheduled media maintenance
- MIGRATION OFFERS OPPORTUNITIES TO:
  - Up-to-data catalog facilities
  - Include low-level data inventory descriptions within catalog inventory file
  - Improve data aggregation to meet current science requirements
  - Compact storage and data transfer through the use of advanced technologies
  - Regeneration of systems and data exercises the data for its health
  - Enables increased on-line access facilities
  - Opens new doors for data access and transfer
- QUALITY CONTROL:
  - Analyze data to ensure credibility during ingest
  - Monitor media performance to ensure reliability over time
  - Maintain log of access activities to build decision history