The NOAA Virtual Data System
Improving Access to and Management of Federal Environmental Data

John Kinsfather
National Oceanic & Atmospheric Administration / National Environmental Satellite, Data & Information Service
National Geophysical Data Center, Boulder, Colorado
Mark McCloy
National Oceanic & Atmospheric Administration / National Environmental Satellite, Data & Information Service
Silver Spring, Maryland

Abstract

The National Oceanic & Atmospheric Administration's (NOAA) National Environmental Satellite, Data & Information Service (NESDIS) operates three national data centers that are geographically dispersed and provide multi-disciplinary environmental data and information to researchers, government planners, business decision makers, lawyers and other professionals, teachers, students, and the general public. At each of these centers there are multiple formats of data stored on a variety of mass storage devices. Data users must cope with numerous access systems related to the specific type of data and primary customers. An individual familiar with these centers who requires only a single type of data can obtain what he/she needs in a reasonable amount of time. It has become essential for NOAA to change the way it manages these data because of (1) an exponential increase in the volume of data archived, (2) a growing demand for the ability to combine multiple types of environmental data, and (3) customer expectation of instant response to data requests. For many years NOAA's management had a vision of a single unified system to more effectively manage these valuable data and to provide faster and easier customer access. In 1996 funds were appropriated to start development of this system.

A new, virtual system is now being implemented to pull these diverse data center operations together. Customers can browse across the distributed data sub-systems and place one order for geophysical, oceanographic, and climatic data, independent of data location, format or storage media. To put these concepts into an operational system with a small budget requires judicious allocation of resources. Numerous storage technologies are being integrated and connected through a scaleable Virtual Private Network (VPN). Other major subsystems of the NOAA Virtual Data System (NVDS) project consist of (1) systems performance monitoring, (2) customer and order management processing, (3) data storage infrastructure, (4) security, (5) metadata management, and (6) customer access mechanisms.

Standards Based Architecture (SBA) is being used in developing the NVDS. Maximum flexibility was a requirement from the outset of the project to be able to quickly adapt to the latest available hardware and software systems and get the biggest payoff from the limited funds available. The system must pull together numerous data sets that vary in size from a few megabytes to tens of terabytes. These data currently exist in thousands of formats and include points, track lines, time series, grids, and images. NOAA management was determined to reduce costs by making use of the maximum amount of COTS (Commercial Off The Shelf) software, implementing standards and minimizing the amount of unique software to be developed and maintained.

User requirements

NOAA's data are as diverse as its user community. These data describe the physical environment of the Earth-Sun system. Users are both internal and external to NOAA. The internal NOAA user community has a strong scientific orientation and consists of researchers, environmental modelers, weather forecasters, planners, and managers. The managers of individual data sets are generally NOAA staff who have a strong scientific background. These data managers are also well versed in applying modern information technologies for effective management of their respective data. NOAA customer service staff are also internal users and these staff generally have limited scientific backgrounds. Many users of NOAA data are also suppliers of data. External users include U.S. and foreign scientists, kindergarten through high school students, universities, industrial companies, governments and the general public. External users cover a wide range of very technical to non-technical people. As shown in Figure 1., users are both on-line and off-line.

NOAA data covers the scientific disciplines of solid Earth geophysics, oceanography, marine geology and geophysics, atmospheric science, climatology, ionospheric physics, space weather, and solar-terrestrial physics. These data contain information about the core of the Earth, the ocean bottom, the water column of the oceans, the land surface, the atmosphere, space environment, and the sun. NOAA has a virtual treasure chest of approximately 1 petabyte of digital data. Terabytes of data are being

collected each day. The data are formatted as points, grids, images, track lines, and time series.

NOAA is faced with the difficult task of making all of these data available to all of these users in an easy, fast, and cost efficient manner. NOAA data are located in many laboratories, offices, and centers throughout the United States. The majority of the data are held at three national data centers: climate data at Asheville, NC; oceanographic data at Silver Spring, MD, and geophysical data at Boulder, CO. The National Climatic Data Center (NCDC), National Oceanographic Data Center (NODC), and the National Geophysical Data Center (NGDC) are referred to collectively as the NOAA National Data Centers (NNDC). A system was needed to link these three major data centers into what would appear as a single system to all users of NOAA data. The goal was to develop a system that would benefit the end user but would also result in operating efficiencies for the agency. If a workable system could be developed among the NOAA data centers then in the future it should also be possible to extend it to other parts of NOAA. Starting in 1996 funds were appropriated to begin development of the NOAA Virtual Data System (NVDS).

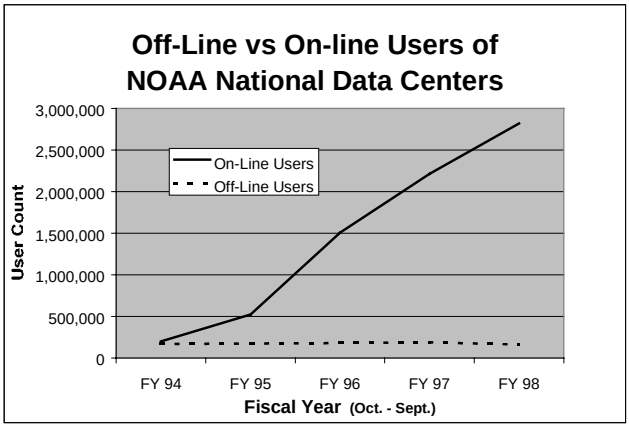

Figure 1. Off-Line vs. on-line users

The NNDC maintain retrospective data and are not responsible for real time operational access to these data. The retrospective nature of the data changes some of the access requirements vs. a real time system. The users are not looking for current weather data to see if a hurricane will hit them but rather they want to look at such things as the tracks for all hurricanes in a given area during the month of August over the last 50 years. This results in a smaller number of users requests than for real time weather data but an expanded requirement to maintain much larger volumes of data in on-line or near-line access mode. Some users of the NNDC data may be able to wait hours or days for their requested data while others may want immediate response to a simple request (What was the ocean surface temperature at a precise location on a specific date?). Often NOAA data users need to have derived environmental information and not just simple raw data. While on-line data access was the driving force in the design of the NVDS, the requirements of NOAA's off-line users of data and derived data products needed to be considered.

During the last five years the NOAA data centers have experienced a drastic change from a totally off-line operation to the current mixed mode of both on-line and off-line operations. NOAA, as well as much of the rest of the world, is going on-line with its data. But, there will be a continuing requirement for NOAA to service off-line users for many years to come. Figure 2 shows the growing demand for data delivered on-line and the continuing demand for off-line data on CD-ROM.

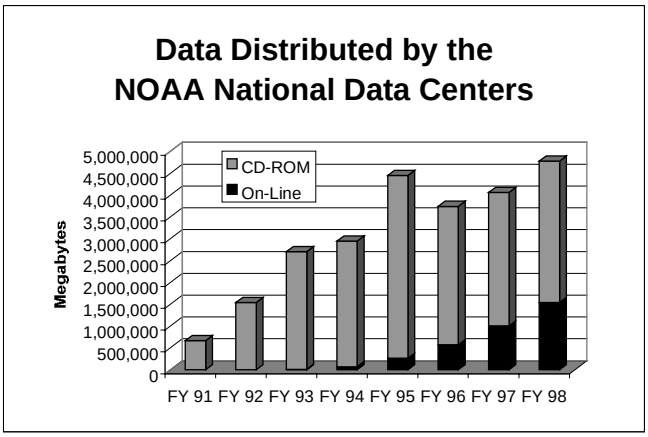

Figure 2. Off-line vs. on-line data distribution

Standards Based Architecture

Standards Based Architecture (SBA) was used to develop the NVDS. The SBA approach follows the planning guidelines from the National Institutes of Standards and Technology (NIST) Application Portability Profile (APP) and the Department of Defense Technical Architecture Framework for Information Management (TAFIM). A series of seven documents evolve from this process to guide subsequent steps in the system design, planning, and implementation. These seven documents are: Architecture Framework [1], Baseline Characterization [2], Target Architecture [3], Opportunity Identification [4], Migration Options [5], Implementation Plan [5], and SBA Assessment. As the system evolves, the content of the documents change from (1) the goals of the new system, (2) an assessment of present capabilities, (3) details of what the new system will do, (4) the projects that must be implemented to complete the new system, and (5) a road map for implementers to follow to achieve the vision of the NVDS. An assessment will be made after completion of the project. Some of the documents have been invaluable to those involved with the project while others have not been used extensively.

The development process has been dynamic. Changes have been made to accommodate new technologies, user requirements, organizational needs, budgets, etc. The goals have remained the same but the path to those goals has changed during the life of the project. The project manager is the Senior Information Officer from the NOAA/NESDIS (National Environmental Satellite, Data and Information Service). A board of directors meets monthly and must approve all major system implementations and budget expenditures. The board consists of the three NOAA Data Center directors and the heads of other NESDIS component offices. The chairperson of the board is the NESDIS Deputy Assistant Administrator for Data and Information Services.

System development process

Many existing legacy data access systems had been developed in NOAA to service the unique requirements of each data set or class of data. In some cases, a data access system had been designed to meet the requirements of diverse classes of users such as physical oceanographers or lawyers. This has resulted in the proliferation of numerous "stovepipe" systems that were based on very real pre-WWW (World Wide Web) user requirements. The challenge of the NVDS was to develop a way to either replace or to combine these legacy systems into one system framework. It did not seem to be feasible for a single system to do everything for all users of all data. But, maybe one system could be developed that would benefit a large percentage of the users of only the most commonly used data sets. Despite
NOAA's large variety and huge volume of data, most requests are for such things as a single observation of air temperature or water temperature for a specific location, date, and time or the present magnetic declination at a geographic location.

The concept of "Build it and they will come" was employed by the NVDS. There was no attempt to mandate the termination of all legacy data access systems. The plan was to build a new system based on modern information technologies and include some of the best components of the old systems. Furthermore, there could not be any interruption in services to customers using the legacy systems. The emphasis was on building an access system that would be attractive (meet user requirements) for both external customers and internal NOAA users. If it could be shown that this new system provided an easier way to maintain data and satisfy customers, then the NOAA data managers would readily embrace it and move their data off the legacy systems. Likewise, if external customers prefer to use the new NVDS access mechanism, then resources can be diverted from continuing the old systems and be used to support the new system. This approach has had some success to date. With future system refinements and iterations, there will be greater acceptance and use of the NVDS. The NVDS will be a dynamic system that continues to evolve to meet the requirements of users, to adapt to

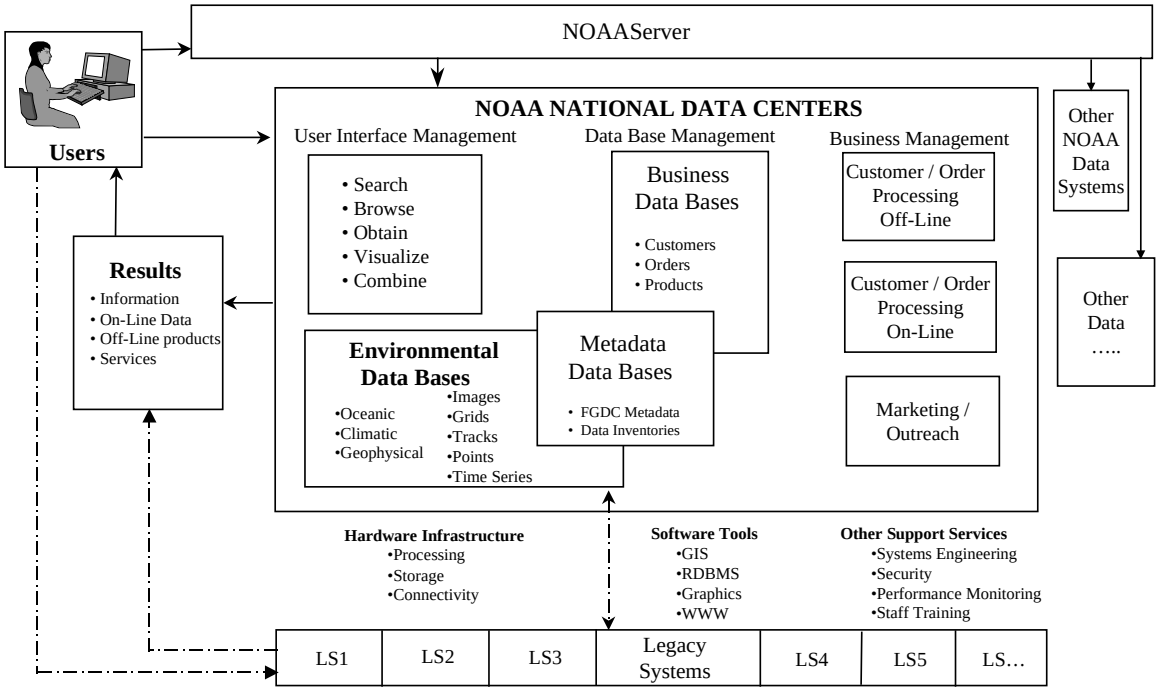

Figure 3. NOAA Virtual Data System components

changing types and volumes of data and to implement the newest information technologies.

Component projects of the NVDS

Initially there were six general component areas for the project:

(1) Planning and management services
(2) Metadata/Data
(3) Data acquisition and processing
(4) Security
(5) Data storage and information infrastructure
(6) User access services

Development of the NVDS is being undertaken by a combination of federal employees and contractors.

It was obvious from the start that the first efforts had to be in rebuilding the basic Information Technology (IT) infrastructure at the NNDC. Continuing budget cuts had resulted in reduced or postponed purchases of new hardware and software for several years. While it was necessary to allocate these funds initially for basic IT infrastructure it also meant that later some of the proposed new and improved components had to be eliminated. Funding these purchases centrally provided a chance to introduce some IT standards among the three data centers. This will reduce support and training costs in the future by having a common knowledge base throughout the organization.

While each Center maintained metadata records in the same standard format, they used a variety of flat ASCII file systems or PC data base systems. The NVDS project provided the opportunity to move all the metadata into a standard data base system and to implement a commercial software system (Blue Angel MetaStar Repository) to manage these metadata. Local copies of metadata are now updated in a Microsoft Access database and a central combined master database is maintained in Oracle. The metadata were also converted from the DIF (Directory Interchange Format) to the FGDC (Federal Geographic Data Committee) format. These metadata now feed into existing NOAA and other government directory systems as well as into the NVDS.

NOAA data centers are required by law to recover the cost of distribution of some of their data products. In general, raw data are made available at no charge and value added products are sold. Prior to the NVDS, each center had its own management information system for maintaining information on customers, orders, billing information and data products. All three centers had plans to implement electronic commerce systems on the Internet. As part of the NVDS, a new management information system, COMPS (Customer Order Management Processing System) was developed. This system will be the core of the NVDS and will tie together the business operations of the NNDCs. A client server system was developed for this purpose. Hindsight indicates that COMPS should have been web-based. It will be converted to a web application in the future. While waiting for the COMPS to be developed, each center implemented an on-line store. These developments were accomplished jointly with the knowledge that the three systems would be merged into a single system when COMPS became operational. A common product catalog was also developed that will be used by all components of the NVDS.

The key component of the NVDS is the user access interface. This interface will provide common search, retrieval and display capabilities for the most popular NOAA data sets. This system uses GIS (Geographic Information System) capabilities to access data stored in an RDBMS and common graphic tools for displaying the data. Most environmental data have a geographic component and these data are readily adaptable to a GIS. The interface system provides transparent access to distributed data in Boulder, Asheville and Silver Spring. This will be the first system to provide generalized data fusion capabilities for NOAA's data. Multiple data sets can be readily accessed, displayed, compared and contrasted and a user can decide to obtain any or all of these data. See Figure 4 for a screen capture of a sample user interface session.

Other components of the NVDS include data outreach/ marketing using off-line methods (mass mailing, conference exhibitions, etc.) and on-line electronic means, computer/network security implementations, and IT training. The training includes vendor specific classes on-site and off-site, computer based training and participation in national IT conferences such as Oracle OpenWorld, COMDEX, InterOp, Sun Java, WWW, etc. The purpose of the training is to build the IT knowledge base in the organization and to do it based on the new set of common software and hardware tools to be used in constructing the NVDS.

One key area that did not get NVDS funding is data acquisition and processing. This was identified as an important component of the NVDS but also as a basic function of the data centers prior to the NVDS. Due to a lack of resources, it was decided that the centers would continue to implement this function completely from internal resources and not from the NVDS. As the implementation progressed, it became necessary to spend more money on on-line systems hardware, software and people. The bottom line for the NVDS is to put the data on-line and make it readily accessible. Only those functions necessary to support that bottom line will be funded.

System development issues

The human factors in building a system are often more difficult to manage than the technology issues. Individual data managers deal with a limited subset of the NOAA clientele and know what is best for "their" users. They do not have the big picture in mind and do not consider possible efficiencies that might be realized on the corporate

level. They have built a career on doing their job in a certain way and don't take kindly to "outsiders" showing or telling them how to do their job more efficiently. Getting a buy-in at all levels of the organization is a difficult but not impossible task. The "carrot" approach has seen better results than the "stick". It is necessary to get input and cooperation from the start of design and throughout the build and implementation phases of a project rather than thrusting a completed system at people and saying, "Here it is, use it!"

NOAA has a long history of interacting with its user community to gain their input in the design of new products, providing new distribution media and developing access systems. The data centers each provide survey forms and solicit input from every customer who receives any off-line data or data product. Each comment received from off-line feedback is responded to individually and most are resolved to the customer's satisfaction. On-line surveys are also employed (see Figure 4.) and many customers provide e-mail input. Each of the data managers has a scientific background (i.e. meteorology, oceanography, geophysics) related to the data they manage. This internal knowledge base and the fact that NOAA is a user as well as a supplier of its own data contribute to a better knowledge of user requirements.

We now operate in a World Wide Web based world and this poses many challenges for developing new systems. The web world also provides many packaged solutions that can be quickly implemented with a minimum of expense and time. The software used to implement the NVDS consists of a large number of COTS and GOTS (Government Off The Shelf) components that are combined with a minimal amount of unique NOAA developed software. The total cost of ownership for any new hardware and software systems being implemented must be known and controlled. Many of the NOAA legacy systems (built with homegrown software by a single individual) will have to be abandoned when the developer retires or leaves the organization. The plan for the NVDS is to eliminate or reduce to a minimum this very undesirable result. In addition to WWW tools, other COTS and GOTS software are being used for RDBMS, GIS, graphics, and format conversions

While the NVDS is trying to reduce the number of access systems for NOAA's data, it is also supporting multiple access paths to the data. There is no reason to cut

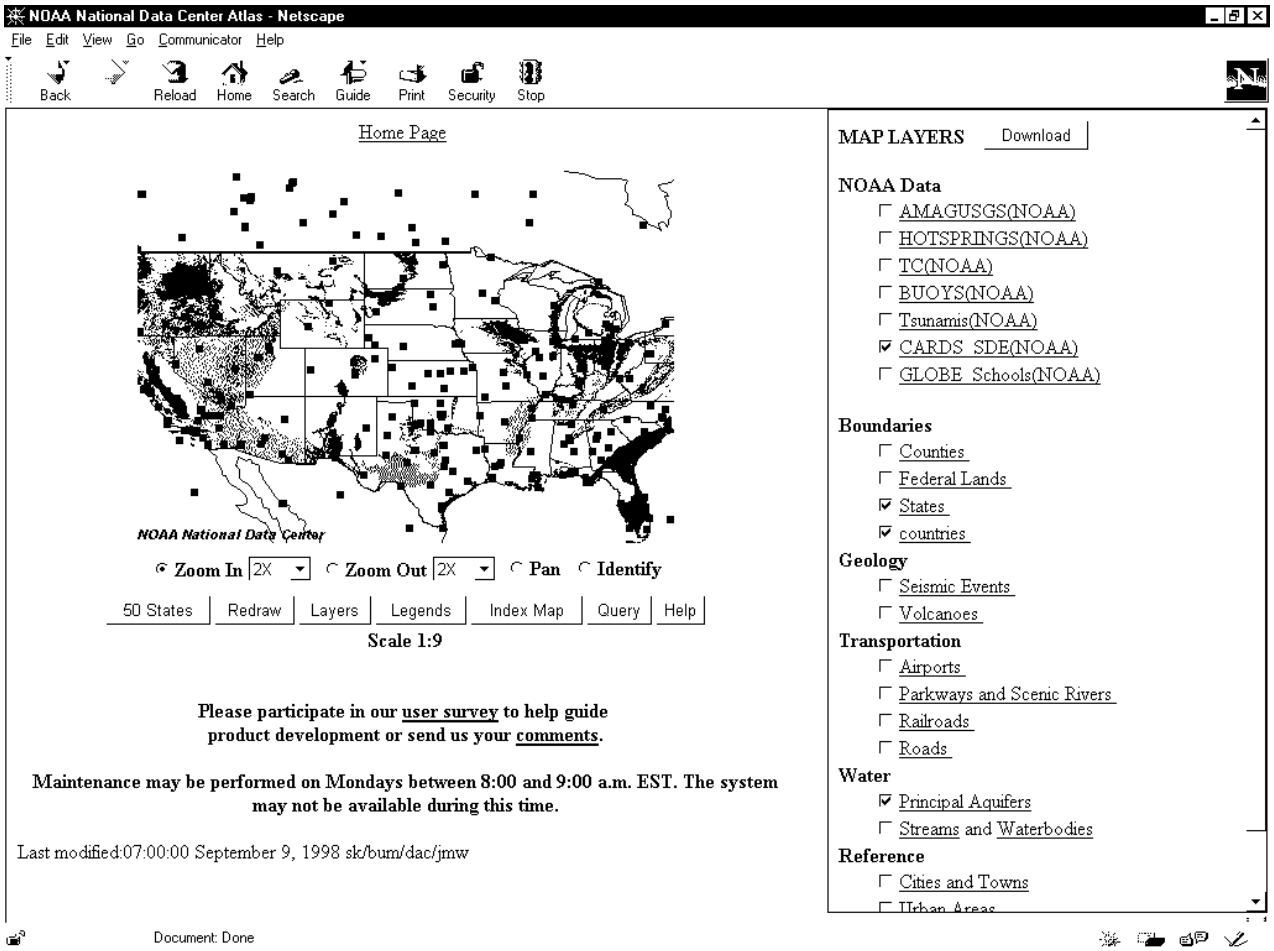

Figure 4. User interface to NOAA data

off existing access methods except where scarce resources needed to be diverted from support of legacy systems to the new way of doing business. Regular customers will see that there is a new and improved way to access data.

Another NOAA data discovery/access system is the NOAAServer. It is a metadata based data discovery tool for many data sets spread across the highly divergent organizational components of NOAA. It is operational across the agency. NOAAServer is an excellent access method for first time users to discover the location of data. The NVDS is building on and integrates with the foundation provided by the NOAAServer. The NVDS proceeds from data search and discovery to data access and delivery from the NNDCs. An on-line user may enter the NNDCs for the first time using the NOAAServer but subsequent searches for and access to data will be expedited by direct use of the NVDS. Once a user finds that he/she has easy and fast access to multiple types of data through a single system, they will be eager to use this new system.

Scaleable architecture is being used in the NVDS to enable integration of both information discovery and automated data handling. Redundant systems ensure that a single point of failure does not exist that could shut down access to all data. A centralized mainframe approach was rejected from the very beginning of the planning phase. Available WWW technologies, a low cost for small but powerful servers, a need for redundancy and an organizational desire to remain geographically distributed were all factors that contributed to this decision.

A distributed system of small to mid-size workstations/servers is being used at each of the NNDCs. The COTS and GOTS software are all very scaleable and run on small Intel processor systems with an NT operating system as well as on larger UNIX hardware. Cost efficient Intel processor systems with a Linux operating system are being used wherever Linux supports the application software. All software is standards based and open systems compliant. Oracle was selected as the NNDC official RDBMS as it is fully supported on the complete range of hardware and by most operating systems. ESRI's Spatial Data Engine (SDE) was selected as the GIS component since it is also fully scaleable across all the hardware platforms, works with most of the operating systems and is fully integrated with Oracle RDBMS. The web server software includes Netscape, Microsoft, Oracle and Apache. Apache is the web server of choice as long as it supports the application software.

Data storage systems include a large range of capacities and media types as shown in Table 1. The NCDC with the largest amount of data uses an IBM Magstar 3590 robotics tape library with Hierarchical Data Storage System (HDSS) software. NGDC and NODC have some data on IBM 3480 tapes. NCDC and NGDC have large volumes of data on 8-mm Exabyte tapes. NODC uses a one-terabyte magneto optical disc jukebox for on-line data access. NGDC provides on-line data from CD-ROM jukeboxes and is starting to write DVD discs and serve DVD data from a jukebox. All centers are making extensive use of RAID systems and non-RAID magnetic disc systems to serve up the majority of their on-line accessible data. The storage systems are based on the volume of data to be handled and in some cases the original media on which the data were collected. Within each Center as well as among all centers there are a variety of locally distributed storage systems. The NVDS goal is to make use of existing storage systems as much as practical, increase on-line storage capacity, and reduce the diversity of systems in the future. The need for numerous types of storage systems of varying capacities will be balanced against the cost of supporting and maintaining a wide variety of systems. Table 1 shows some of the storage systems and media in the NNDC.

	NCDC	**NGDC**	**NODC**
Data sets	700	400	200
Archive tbyte	640.0	10.5	1.2
Tape media 1	*3590*	*3480*	*3480*
Media #	2,000	5,200	200
Tape media 2	*3480*	*8 mm*	*8 mm*
Media #	300,000	6,750	200
Tape media 3	*8 mm*		*9-track*
Media #	40,000		2,000
On-line tbyte RAID	1.0	0.1	0.3
On-line tbyte non-RAID	1.0	0.5	0.28
Robotics system(s)	*Magstar 3590*	*DVD-ROM CD-ROM*	*Magneto Optical*
Near-Line tbyte	4.0	0.5	2.2
Staff	270	123	104

Table 1. NOAA National Data Centers size parameters

In order to tie together three geographically dispersed data centers into one logical system, efficient networking is needed. NOAA relies on the Internet for much of its network requirements. The Internet can provide some cost efficiencies but may not always provide the bandwidth and response time required by data users. Part of the IT infrastructure improvement at the NNDC was the development of high speed (100+ megabits/second) switched local area networks. A frame relay network among the three centers was also implemented. The frame relay wide area network was primarily implemented to support the common Management Information System for the NVDS but plans are in place to increase the bandwidth and use this network for environmental data transfers.

Lessons learned in system development

When developing a system among multiple organizations or among multiple semi-autonomous groups

within one organization, the problems are generally more business and administrative than they are technological. Implementing new information technologies is easy, working with people from diverse organizations can be difficult. A vision and common goals should be communicated often to everyone involved in a project. A buy-in by everyone involved is needed from the very beginning of a project in order to have a feeling of ownership. There is never one solution that will fit all requirements, fit all data or fit all users. Stovepipe legacy systems should be replaced where appropriate but one should not assume that all legacy systems need to be eliminated. Try to build a new system on top of the best components of existing systems, where appropriate. GIS access capabilities are appropriate for many types of data but not necessarily for all data.

Finding commonalties to build on is one of the keys to a project's success. It is relatively easy to see that most environmental data users want on-line access to data using the Internet. But improving on-line access does not mean totally eliminating traditional off-line access to data. Efforts should be concentrated on getting the highest cost/benefit ratio. Everyone agrees that customers expect easy and fast access to as much data as possible, but is that a reasonable capability to provide for all of an organization's data? Try to create a system that will meet 100% of your users' requirements but be aware that perhaps the 90% solution can be accomplished much faster and at half the cost. The 100% solution might never be completed. Team building, education and training are effective tools that should be employed for any large project.

The major funding for development of the NOAA Virtual Data System will terminate in September 1999. There will be a smaller level of ongoing funding for basic support and maintenance. Many, but not all of the system concepts and components have been implemented. The user interface access system will be operational in the middle of 1999. The success of the NVDS will be determined by user satisfaction. Will it be easy for customers to use this interface and get back the results they want?

The NVDS is a dynamic system and will continue to evolve. New data management and data access capabilities will need to be implemented. Large volumes of additional data need to be added to existing mass storage systems. New storage systems with larger capacities will also be required. NOAA is optimistic but knows that there is a lot of work yet to be done. In order to continue and expand on the development work of the NVDS, plans are now being prepared for "NVDS2."

References

1. Architecture Framework Document for the NOAA Virtual Data Center (VDC), West Virginia High Technology Consortium Foundation Contract No. WVHTC-F-S95-1019, February 14, 1996

2. Baseline Characterization Document for the NOAA Virtual Data Center (VDC), West Virginia High Technology Consortium Foundation Contract No. WVHTC-F-S95-1019, February 14, 1996

3. Target Architecture Document for the NOAA Virtual Data System (NVDS), March 28, 1997

4. Opportunity Identification Document for the NOAA Virtual Data System (NVDS), August 19, 1997

5. Migration Options & Implementation Plan Document for the NOAA Virtual Data System (NVDS), Marada Corp. Contract Number: 50-SPNA-7-00001, Task Order Number: 56-SPNA-7-0001, November 7, 1997