



National Snow and Ice Data Center
Supporting Cryospheric Research Since 1976

Challenges in Long-Term Data Stewardship

Ruth Duerr
University of Colorado at Boulder
Boulder CO 80309-0449
+1-303-735-0136
rduerr@nsidc.org

NASA/IEEE MSST 2004

12th NASA Goddard/21st IEEE Conference on
Mass Storage Systems & Technologies
The Inn and Conference Center
University of Maryland University College
Adelphi MD USA
April 13-16, 2004





National Snow and Ice Data Center

Supporting Cryospheric Research Since 1976



I'd like to thank my coauthors:
Mark A. Parsons, Melinda
Marquis, Rudy Dichtl, Teresa
Mullins

Thanks also to Jenny Jenkins and Wendy Thoreaux
for their review and input

Presentation Overview

- A Brief History of Scientific Data Stewardship
- Scientific Data Stewardship Defined
- Differences Between Digital Preservation in the Library and Science Data Contexts
- Q&A
- Data and Metadata Challenges
- Scientific Stewardship Related Challenges
- Q&A

A Brief History of Science Data Stewardship

- The distant past
 - Historically scientific data were recorded in notebooks, logs or in maps
 - With luck a library or archive would collect and preserve these
 - Finding and accessing data were difficult

A Brief History of Science Data Stewardship (cont.)

- In recent centuries
 - The establishment and growth of academic and public libraries improved the situation
 - Librarians became data stewards, developing cataloging, indexing, preservation and accessing schemes
 - Data were still analog

A Brief History of Science Data Stewardship (cont.)

- 1957 - World Data Centers (WDC)
 - Established during the International Geophysical Year
 - Focus on preservation and distribution of raw data
 - Organized by discipline
 - Data were still analog

A Brief History of Science Data Stewardship (cont.)

- After the 1960's
 - Discipline-specific data centers proliferate
 - Federal government
 - A total of 9 national data centers were established in the US
 - Sponsored by NOAA, NASA, USGS, DOE
 - Focus on archival and distribution of data
 - Local and state governments
 - Universities
 - Commercial Entities

A Brief History of Science Data Stewardship (cont.)

- 1990's Earth Observing System (EOS)
 - A large system of remote sensing instruments and data systems
 - Distributed Active Archive Centers
 - 8 discipline specific centers designated by NASA
 - Typically co-located with established data centers
 - Focus on archive and distribution during most active part of the data life cycle
 - Provides a web-based interface to simultaneously search and access data from all the DAACs as well as data centers scattered around the world

A Brief History of Science Data Stewardship (cont.)

- What is important to note about EOS and the DAACs is that they were arguably the functional beginning of a new data management model:
 - Geographically distributed data archival
 - Centralized search and order

A Brief History of Science Data Stewardship (cont.)

- Trends that continue today
 - Centralized access to decentralized data
 - Inadequate planning for long term data preservation

A Brief History of Science Data Stewardship (cont.)

- Role of the World Wide Web in decentralizing data storage
 - Search engines
 - Theoretically assist with locating data
 - Rarely provide sufficient information about the utility of the data
 - Data reliability and integrity are not verifiable
 - The content of the web is ephemeral
 - Users expect ready access to data

A Brief History of Science Data Stewardship (cont.)

- Role of private records management companies
 - Cost/benefits analysis
 - International policies regarding access to data

Long-Term Stewardship Defined

Within the data management field the phrase “long-term” typically is defined as:

“A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository.”

OAIS Reference Model

Long-Term Stewardship Defined

- Notions of stewardship are less well defined
- Three relevant definitions:

“the person or group that manages the development, approval, and use of data within a specified functional area, ensuring that it can be used to satisfy data requirements throughout the organization”

DOD Directive 8320.1-M.1

Long-Term Stewardship Defined (cont.)

- Three relevant definitions (continued)

Long-term archiving needs to be a “continuing program for preservation and responsive supply of reliable and comprehensive data, products, and information ... for use in building new knowledge to guide public policy and business decisions”

Global Change Science Requirements for Long-Term Archiving

Long-Term Stewardship Defined (cont.)

- Three relevant definitions (continued)

“maintaining the scientific integrity and long term utility of scientific records”

NOAA/NESDIS, 2003

Long-Term Stewardship Defined (cont.)

- These definitions associate the notion of science stewardship with two concepts
 - Data preservation
 - Access or use in the future

Why preserve data?

- To ensure its utility for users in the future. Some examples include:
 - To allow combination with historical data to assess change over time
 - To allow future development of new or improved products
 - For use of data in ways that were not originally anticipated
 - To permit replication of scientific results

Preservation in the Historical Library Context

- Library patrons expect to experience the material preserved
- Library patrons do not expect to be able to transform the accessed materials
- Library patrons are typically less concerned with how the original object was created
- Libraries are therefore concerned more with issues such as whether and how to preserve the “look and feel” of an object

Preservation in the Science Data Context

- Users expect to be able to manipulate the data retrieved
- Users even expect to receive data that has been transformed during the process of extracting it from the archive
- Scientists also need to understand how the data were created
- Data archives are therefore more concerned with preserving the bits and their meaning as well as information about how the data were created

Information About the Data that Must be Preserved

- “Instrument/sensor characteristics including pre-flight or pre-operational performance measurements (e.g., spectral response, noise characteristics)
- Instrument/sensor calibration data and method
- Processing algorithms and their scientific basis, including complete description of any sample or mapping algorithm used in the creation of the product (e.g., contained in peer reviewed papers, in some cases supplemented by thematic information introducing the data set or product to scientists unfamiliar with it)
- Complete information on any ancillary data or other data sets used in generation or calibration of the data set or derived product”

Global Change Science Requirements

Information About the Data that Must be Preserved (Cont.)

- “Processing history including version of processing source code corresponding to versions of the data set or derived product
- Quality assessment information
- Validation record, including identification of validation data sets
- Data structure and format, with definition of all parameters and fields
- In the case of earth-based data, station location and any changes in location, instrumentation, controlling agency, surrounding land use and other factors which could influence the long-term record”

Global Change Science Requirements

Information About the Data that Must be Preserved (cont.)

- “A bibliography of pertinent Technical Notes and articles, including refereed publications reporting on research using the data set
- Information received back from users of the data set or product”

Global Change Science Requirements

Break for Questions

Presentation Overview

- A Brief History of Scientific Data Stewardship
- Scientific Data Stewardship Defined
- Differences Between Digital Preservation in the Library and Science Data Contexts
- Q&A
- Data and Metadata Challenges
- Scientific Stewardship Related Challenges
- Q&A

Data and Metadata Challenges

- Standards
- Preservation vs Access
- Separation Issues
- Data Security and Integrity
- Long-Term Preservation and Technology Refresh
- Size Does Count!



Standards

- The EOS Core System (ECS) experience
 - Community-based standards work best
 - Standards profiles that support a particular community may be needed
- Plethora of types of standards - for example
 - Data format
 - Metadata types, content, and format
 - Documentation format and content

Format Standards ★

- The challenges of preserving information stored in proprietary formats are well known
- Increasingly this is an issue for ancillary information about the data as well as for upper level data products
- Even non-proprietary standards change over time

Format Standards (continued)

- Proposed solutions
 - Digital format archives 
 - Archival in a technology independent representation (e.g., Universal Data Format) 
 - Keep the archive simple and format the data for the user on the fly

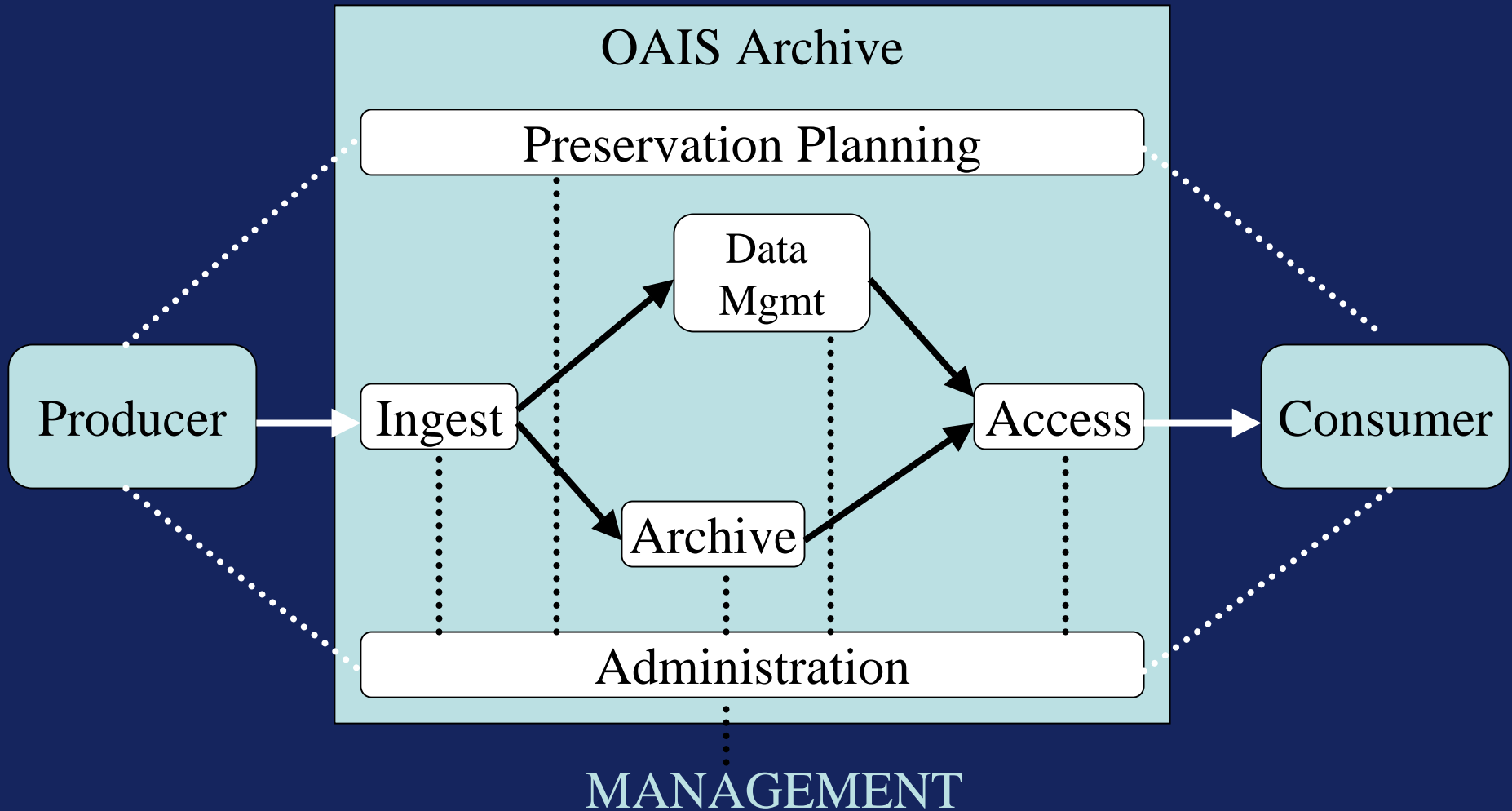
The OAIS Reference Model

- A CCSDS and ISO standard that describes data preservation concepts such as:
 - Responsibilities of an archive
 - Functional model describing how to preserve information and make it available to users
 - Information model describing what ancillary information is needed to ensure that future users understand and can use the information preserved
 - A common set of terminology that can be used to describe the above

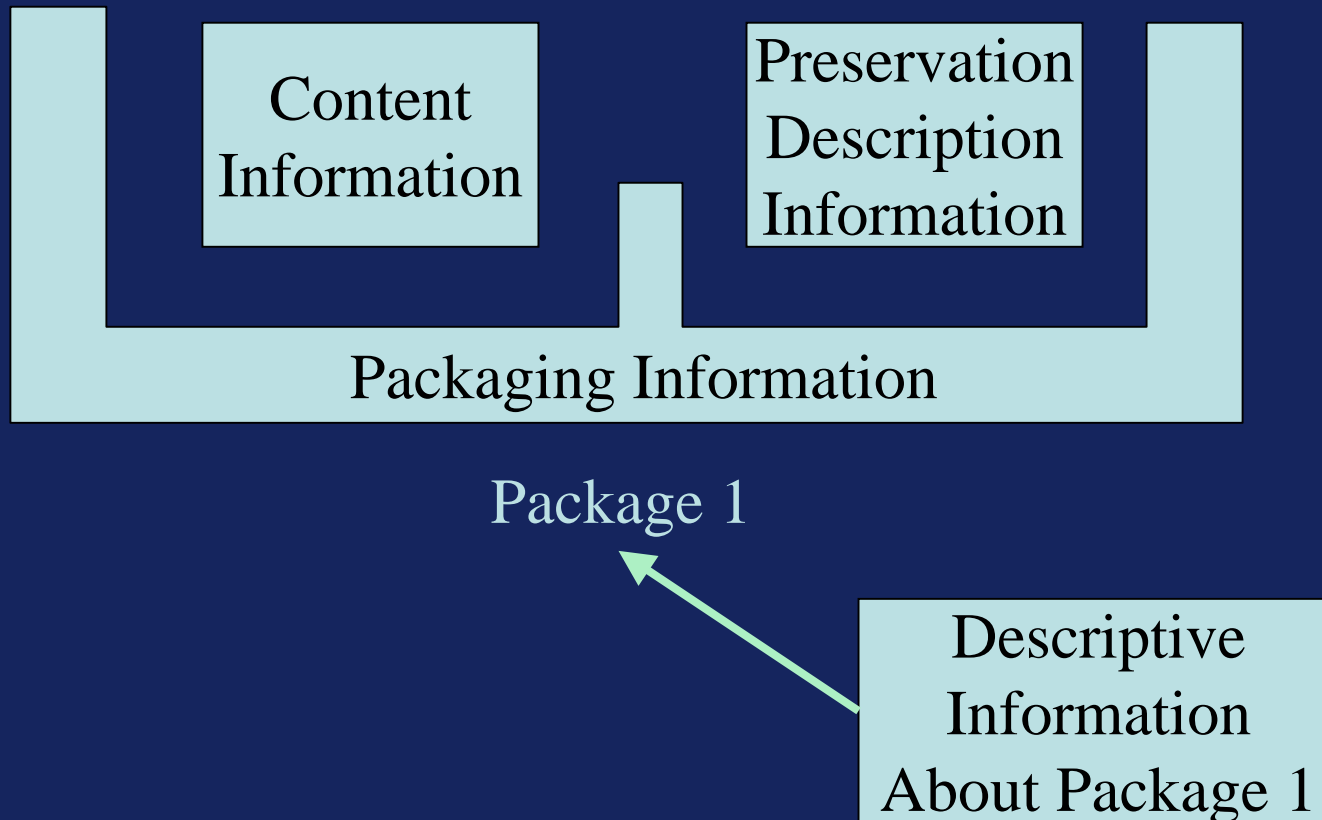
OAIS Archive Responsibilities

- Negotiate with information providers to receive and obtain sufficient rights to appropriate information to ensure long-term preservation★
- Designate a community which should be able to understand the information preserved
- Ensure that the information is independently understandable to that community★
- Document procedures and policies regarding data preservation and access★
- Make the information available

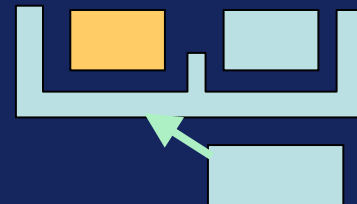
OAIS Functional Model



OAIS Information Model



OAIS Information Model - Content Info.



- Data Object - the information to be preserved
- Representational Information - allows a user to understand the data
 - Structure (e.g., flat binary file, ASCII table, net-CDF file, HDF, etc.)
 - Content (e.g., a table of station IDs, dates, latitude, longitude, incidence angle, brightness temperature)

OAIS Info. Model - Preservation Description



- Provenance - documents the history of the object
- Reference - documents object identifiers and their generation mechanisms
- Fixity - documents methods used to ensure there are no undocumented changes
- Context - the relationship of the object to its environment

OAIS Preservation Description - Provenance ★

- Information about the pedigree/history of the data
 - Where did it come from and where has it been since?
 - Who created it?
 - How was it created; what algorithms, algorithm versions, ancillary and calibration data sets were used?
 - What other data were used to validate these data?
 - What changes have taken place since these data were originally created?

OAIS Preservation Description - Reference

- Persistent, unambiguous identifiers ★
- Aliases commonly in use
- A description of the rules (if any) for creating the identifier


OAIS Preservation Description - Fixity ★

- Authentication information
 - Descriptions of the mechanisms used to ensure that the data has not been changed in an undocumented way
 - Authentication keys
 - Fixity information is uncommon

OAIS Preservation Description - Context

- Information context
 - Why were the data created?
 - How do these data relate to other data?

Beyond the OAIS Reference Model

- The OAIS Reference Model was not intended to be a design or implementation level standard
- The document discusses a wide variety of implementation level standards that could be developed 
- Several organizations have either defined their own preservation metadata format or are working on doing so

The RLG/OCLC Metadata Framework

- The Online Computer Library Center and the Research Libraries Group sponsored development of a preservation metadata framework for digital objects based on the OAIS model
- The framework
 - defines schema elements for preservation metadata
 - does not specify implementation level details
 - allows expansion of lowest level elements

Other Preservation Metadata Activities

- The OCLC PREMIS subgroup is working on defining a set of core attributes and implementation strategies
- The Dublin Core Metadata Initiative's Preservation Working Group is working on a charter which will include investigation of the need for domain specific preservation metadata schemas
- Recently the CODATA group has started to look at whether there is a need to define preservation metadata schema for science data

Content Standard for Digital Geospatial Metadata

- Established by the Federal Geographic Data Committee
- All federally funded programs that involve geospatial data are required to adhere to this standard
- Purpose is to allow users to find geospatial data, assess its utility for their purposes and to access the data
- Has some overlap with the OAIS reference model

ISO 19115

- More or less the international version of the FGDC standard
- A “cross-walk” between the FGDC and ISO standards exists
- Is a content standard, not an implementation standard

- Consensus seems to be building that whatever the schema, XML should be the implementation standard for metadata
 - ISO Technical Committee 211 is developing a UML implementation standard for ISO 19115 that will include an associated XML schema
 - NRC report on “Government Data Centers: Meeting Increased Demands” also recommends XML

Preservation vs Access

- Science data users want data in easy to use forms
 - May wish to receive the data in a specified format
 - May wish to obtain only a particular subset of the data
 - May wish to have the data re-gridded or re-projected

Preservation vs Access - Implications

- Science data archives may need interfaces supporting many different data access formats, grid types and projections
- Access formats are likely to change over time

Preservation vs Access - Strategies

- Separate preservation and access storage
- Storage as a simple technology-independent stream of bytes, with adequate “representation information” with “format on the fly” access capabilities★
- Storage in a database✶★

Separation Issues ★

- Storing data and their associated metadata separately increases the risk that they will become detached
 - May impede utility
 - May result in misuse
- Separation can occur even if simple techniques such as 'tar' are used
- Embedding the metadata within the data can solve this but raises other issues

Separation Issues (continued)

- The situation is exacerbated when the data and metadata start out geographically separated
 - Even preservation of the data can be at risk in this situation

Separation Issues - Brokered Products

- NSIDC is often tasked with creating metadata and advertising products held elsewhere
- When users request data they are referred to the external site holding the data
- Simply maintaining the links to these external sites is a challenge

Separation Issues - the CAPS Example

- NSIDC collaborating with the International Permafrost Association released a CD titled Circumpolar Active-Layer Permafrost System in 1998
 - A major milestone of the Global Geocryological Data (GGD) system
 - The CD held 56 data sets and references to about 100 more held at other “nodes” of the GGD system

Separation Issues - The CAPS Example (cont.)

- Unfortunately funding for the GGD stopped in 1998
- In 2002 a new initiative started - creating an updated version of the CD was high on the to do list
- Dozens of the original “brokered” products are no longer readily available

Data Security and Integrity

- Ensuring the integrity of the data involves at least three components
 - The data must demonstrate scientific integrity
 - The data must not have been altered since creation
 - Adequate preservation practices exist

Data Security and Integrity - Scientific Integrity

- Notions of scientific integrity are rooted in the concept of the scientific method
 - Experiments must be repeatable
 - Results should be published in peer-reviewed literature
 - Data and information used must be specifically acknowledged and accessible

Acknowledging Data and Information

- Traditionally data have been published in journals or monographs that could be specifically cited
- Currently methods vary by author
 - Simple acknowledgement of the data source in the paper
 - Often difficult to trace especially over time
 - Often imprecise
 - Sometimes do not acknowledge the true data source

Acknowledging Data and Information (cont.)

- Currently methods vary by author (cont.)
 - Citation of an article published by the data provider that describes the data set and its collection
 - May not exist in the peer-reviewed literature
 - May only describe a portion of the data set
 - May not be relevant to this new application of the data
 - May not allow readers to acquire the data and even if it does the information may degrade over time

Acknowledging Data and Information (cont.)

- Currently methods vary by author (cont.)
 - Use of data citations
 - What is a data citation?
 - Typically the “author” is the data provider or person who invested intellectual effort into creating the data set
 - Typically the “publisher” is the archive that distributed the data
 - The publication date is used to distinguish different versions of related data sets

Acknowledging Data and Information (cont.)

- Currently methods vary by author (cont.)
 - Use of data citations (continued)
 - Publisher information may degrade over time
- With the rise of “electronic journals” the concept of including the data within the publication has been informally discussed
 - The electronic journal becomes a science data archive with all the attendant challenges

Ensuring the Data Received was as Expected

- The “fixity” issue from the OAIS reference model
- Often this is described as a problem that is solved - not true!
 - Using message digest algorithms such as MD5 to ensure that the data sent is the data received
 - Then using digital signature technologies to ensure that the data came from a reputable source

Ensuring the Data Received was as Expected

- Issues
 - Resources required
 - Algorithm/mechanism stability over time
 - Based on the reputation of the data source

Trusting the Data Source

- The user must be able to trust that the preservation practices of the source are adequate. For example:
 - Archive media are routinely verified and refreshed
 - Facilities are secure
 - Processes to verify and ensure the fixity of the data are operational
 - Adequate mechanisms exist to ensure data can be recovered in case of emergency
 - Disaster recovery plans and procedures are in place

Trusting the Data Source (continued)






- RLG/OCLC Working Group on Digital Archive Attributes suggests that processes for certifying digital repositories be put in place
- It has been suggested that folks with administrative access to data and metadata be subject to “strong proofs of identity”

Long-Term Preservation & Technology Refresh

“digital objects require constant and perpetual maintenance,
and
they depend on elaborate systems of hardware, software, data and information models, and standards that are upgraded or replaced every few years”★

NSF and Library of Congress, August 2003

Long-Term Preservation & Technology Refresh

- Three proposed solutions
 - Normalization - Conversion to a few “technology independent” standard formats on ingest  
 - Migration - transferring data to new technologies before the old become obsolete 
 - Emulation - recreating the original environment on current technologies  

Size Does Count!

- Generally there are many more small data sets than large data sets
- Most of the collection level metadata creation resources are needed for these small data sets
- Collection level metadata needs are more or less independent of data set size
- How can automated metadata generation tools mitigate these resource needs? ✨★

Scientific Stewardship Challenges

- Maintaining science understanding over time
- Decisions, Decisions, Decisions -
Deciding what data to acquire and retain
- Upfront planning

Maintaining Science Understanding Over Time

Scientists need to be involved with the data

- Maintain data integrity over time
- Avoid misapplications of data
- Address known limitations of data
- Include information on data harmonization and improvements

Decisions, Decisions, Decisions

- Deciding what data to retain - the problem
 - Impractical to retain all data for all time
 - Effective business models for cost/benefits of long-term data archive do not exist★
 - History shows us that many data sets have unanticipated future applications★
 - In order for results to be reproducible, the data used must remain accessible

Decisions, Decisions, Decisions

- Preservation Options ★
 - Preserve all levels of the data for all time
 - Preserve the lowest level of the data along with the algorithms to create higher level products
 - Preserve only the processed products
 - Preserve only products that have been requested

Upfront Planning

- The best time to start thinking about data stewardship is at the very beginning
 - Doing otherwise puts the data at risk
 - Doing so can increase the quality and availability of not only the metadata but also the data

An Example - The CLP Experience

- NSIDC was involved from the start
- NSIDC management folks were in the field
 - Interviews and follow up with the investigators
 - Manual and automated QC of the data collected each day
- Resulted in better QC documentation and higher-quality data



Summary

- Preservation of digital data presents many challenges
- Some of which are exacerbated when data is distributed
- Technology can be used to mitigate many of these challenges; however,
- People, especially scientists, need to be involved to maintain the scientific integrity of these data over time

Break for Questions

For More Information

- About NSIDC in general
 - <http://nsidc.org>
 - nsidc@nsidc.org
- About data management or archiving at NSIDC
 - rduerr@nsidc.org
 - (303) 735-0136