National Snow and Ice Data Center

Supporting Cryospheric Research Since 1976

Challenges of a Small Archive

Ruth Duerr

NSIDC/CIRES/University of Colorado - Boulder

1540 30th St., Boulder, CO, 80903-0449

Phone: +1-303-735-0136 FAX: +1-303-492-2468

E-mail: rduerr@nsidc.org

Presented at the THIC Meeting at the National Center for Atmospheric Research, 1850 Table Mesa Drive, Boulder CO 80305-5602

June 29-30, 2004

The Premier Advanced Recording Technology Forum

THIC Inc.

Overview

- NSIDC Overview
- The Digital Data Preservation Challenge
- NSIDC Systems - Now & Near-Term

Institutional Relationships

- Part of the University of Colorado
- Within the CU Cooperative Institute for Research in Environmental Sciences (CIRES) Cryospheric and Polar Processes Division
- Chartered by NOAA's National Environmental Satellite, Data, and Information Service. Affiliated with the NOAA National Geophysical Data Center (NGDC)
- Funded by NASA, NOAA, NSF, and others at the program level
- Part of the World Data Center system

Major Programs

NASA Distributed Active Archive Center

NOAA at NSIDC and WDC for Glaciology, Boulder

NSF Arctic System Science Data Coordination Center

NSF U.S. Antarctic Data Coordination Center and Antarctic Glaciological Data Center

IARC Frozen Ground Data Center

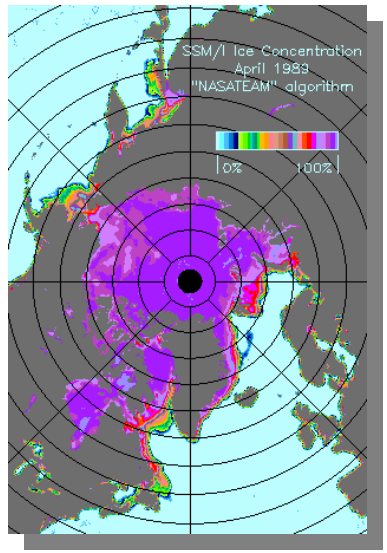
Our mission

To make fundamental contributions to cryospheric science and excel in managing data and disseminating information in order to advance understanding of the Earth system.

Data Management and Distribution

Outreach

International Data Activities



Research

Data Holdings

- ~600 archived digital data sets
  - ~400 are publicly available
  - Range in size from a few KB to many TB
  - From a single file to millions of inter-related files
  - Many of the smaller data sets are digital transcriptions of analog data, some of which date back hundreds of years
  - Most of the larger data sets are from satellite remote sensing instruments
- Archives grow by 2-4 TB/month (~75 TB currently)
- 1-4 TB distributed each month

Why do we preserve science data?

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

Science Archives & Libraries

Library Users:

- expect to experience the material preserved
- do not expect to be able to transform the accessed materials
- are typically less concerned with how the original object was created

Libraries are concerned more with issues such as whether and how to preserve the “look and feel” of an object

Science Archive Users:

- expect to be able to manipulate the data retrieved
- even expect to receive data that has been transformed on the way
- need to understand how the data were created

Science archives are more concerned with preserving the bits, their meaning, as well as information about how the data were created

The Digital Preservation Challenge

"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

Constant and Perpetual Maintenance?

- Improvements in scientific understanding necessitate occasional reprocessing
  - Need to manage multiple versions of the product
  - Need to make scientists working with the data aware of the updated products
  - Need to maintain lower level products

Constant and Perpetual Maintenance? (continued)

- User communities drift over time and different communities likely have different needs
  - E.g., remote sensing data communities
    - Science team members responsible for calibrating/validating the data
    - Scientists working in the field
    - General science community
    - General public
    - Decision makers
- Even language drifts over the long term

Standards

- Archive Standards
  - OAIS Reference Model
  - Global Change Science Requirements for Long-term Archiving
- Metadata Standards
  - Content
    - NASA DIF moving to FGDC
    - FGDC moving to ISO 19115
    - OAIS derived preservation metadata standards

Standards (continued)

- Metadata Standards (continued)
  - Format
    - ASCII text
    - ECS .met files
    - XML
- Data format standards
  - Flat binary files
  - ASCII
  - HDF-EOS
  - Shape files

Current Systems

- ECS System - Primarily SGI and Sun based
  - Archive
    - STK Powderhorn w. 9940a drives (~300 TB capacity)
    - ADIC AMASS software
  - Ingest
    - All electronic mostly using NASA ECS interfaces
  - Distribution
    - DVD-R, CD-R, DLT, 8mm tape
    - ftp pull/push

Current Systems

- Non-ECS Systems - Primarily SGI and Sun based
  - Archive
    - STK Timberwolf w DLT drives
    - 4 TB of fiber attached RAID
  - Ingest
    - Ftp pull/push, fastcopy, email
    - 8mm tape; DLT, CD, DVD
    - External disk
  - Distribution
    - DLT, 8mm tape, CD, DVD+R
    - Ftp pull

Near-Term Trends at NSIDC

- Increasing amounts of RAID
- Increasing dominance of Linux
- Need to determine what tape path to take (I.e., stick with DLT or switch to LTO)
- Testing viability of replacing the Timberwolf by a COPAN MAID system

For More Information

- About NSIDC in general
  - http://nsidc.org
  - nsidc@nsidc.org
- About data management or archiving at NSIDC
  - rduerr@nsidc.org
  - +1-303-735-0136