Data Management and Digital Preservation for Arctic Science: CADIS and Chronopolis

Eric Nienhouse
Michael Burek
Don Middleton, Nate Wilhelmi
Hannah Wilcox, Danielle Harper
NCAR
Outline

Arctic Science Meets Digital Preservation

- What is digital preservation?
- SGF Case Study: CADIS Arctic Data Portal
- Digital preservation pilot using Chronopolis
- Lessons learned
Digital Preservation

Ensuring long term usability of precious collections.

- Certain research data are unique and cannot be replaced.
- Digital information loss is a real problem (NSF & LOC)
- Legal stewardship requirements for publically funded activities.
- Data preservation is not data backup.
- Preservation requires active data management and curation.
Case Study: CADIS

Arctic Data Discovery and Access

- NSF funded project to catalog Arctic data collections
- Collaboration between NCAR, Unidata and NSIDC
- In support of the Third Int’l Polar Year (2007 – 2008)
- IPY efforts (1882, 1931) generate irreplaceable data collections
- Historic IPY data collections (1882, 1931) scattered and lost
- Opportunity to develop a long-term collection from IPY3
Case Study: CADIS

NSF Vision for Arctic Observing Network

- AON would be “a system of monitoring that will significantly advance our observations of Arctic environmental conditions”
- Implied considerable data integration and data stewardship
- In reality AON is a number of field projects with individual needs
- Diverse platforms, metadata and data formats
- Few systems and interfaces for interoperability
Case Study: CADIS

Initial Steps Toward NSF Vision

- Need to respond quickly to AON project needs
- Establish and interview a representative user group
- Developed CADIS Metadata Profile and use cases
- Leverage existing SGF data management tools
- Provide guidance and support for project investigators
- Develop data conversion tools and documented processes
CADIS Data Portal

SGF Gateway

- Stand Alone Gateway Node
- Data Publication Web Interface
- Access Control and User Management
- Faceted and GeoSpatial Search, Collection Browsing
- OAI Metadata Exchange OAI-DIF (GCMD, Damocles)
- Data Access and Download (HTTP, wget)
- Metrics capture and funding agency reporting
- SGF Version 1.3 RC1 deployed and in testing phase
CADIS Metadata Profile and Editor

- Editor and profile key success factors
- Based on WMO IPY-DIS profile
- Compatible with GCMD, FGDC, ISO
- Profile driven UI with validation
- NASA GCMD vocabulary
- Curation by metadata team
- Integrated end user help documentation
CUDA IS Now: Content Summary

- 64 AON projects
- 89 Lead and collaborating investigators
- 355 archived data collections
- 110 brokered data collections
- 465 complete and reviewed metadata records
Chronopolis Pilot

Builds on Sound Metadata Foundation

- CADIS Archive Requires Digital Preservation
- Precious set of data collections
- NSF data management mandate
- Ideal collection size for preservation pilot
- NCAR primary partner with Chronopolis consortium
- NCAR Chronopolis node proximity and network (e.g. TeraGrid)
Chronopolis: Quick Facts

A Digital Preservation Network Funded by The Library of Congress

- Designed to leverage capabilities at multiple institutions
- All partners provide network and dark archive storage services
- Emphasizes heterogeneous and redundant data storage systems
- Has a current storage capacity of 300 TB (100 TB at 3 nodes)
- Geographically distributed copies of data (NCAR, UMD, SDSC)
- Includes detailed monitoring and monthly auditing of all data
Chronopolis: Technology

Integration Effort of Available Components

- Built upon SRB/iRODS storage system
- ACE Audit Control Environment ensures integrity
- Bagit manifest for collection packaging
- Captures fixity data at source
- Grid security via proxy certificates
- System monitoring web interface
Chronopolis: System Monitoring

Web Monitor Interface

- Quick system overview
- Aggregates node state
- Spring Application Framework
- YUI, AJAX
CADIS: Lessons Learned

Community Engagement Had the Strongest Effect

- Investment in metadata profile really paid off.
- Community surveys, demos and use cases critical.
- Well defined and responsive user support team success factor.
- Data curation and user support costs were double expectations.
- Achieving vision for Arctic Observing Network will take time.
Thank You!

In particular:

- Michael Burek, NCAR
- NSIDC, NCAR EOL, Unidata, SDSC, UCSD Library, Library of Congress, UMICS, and others.
Questions?