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

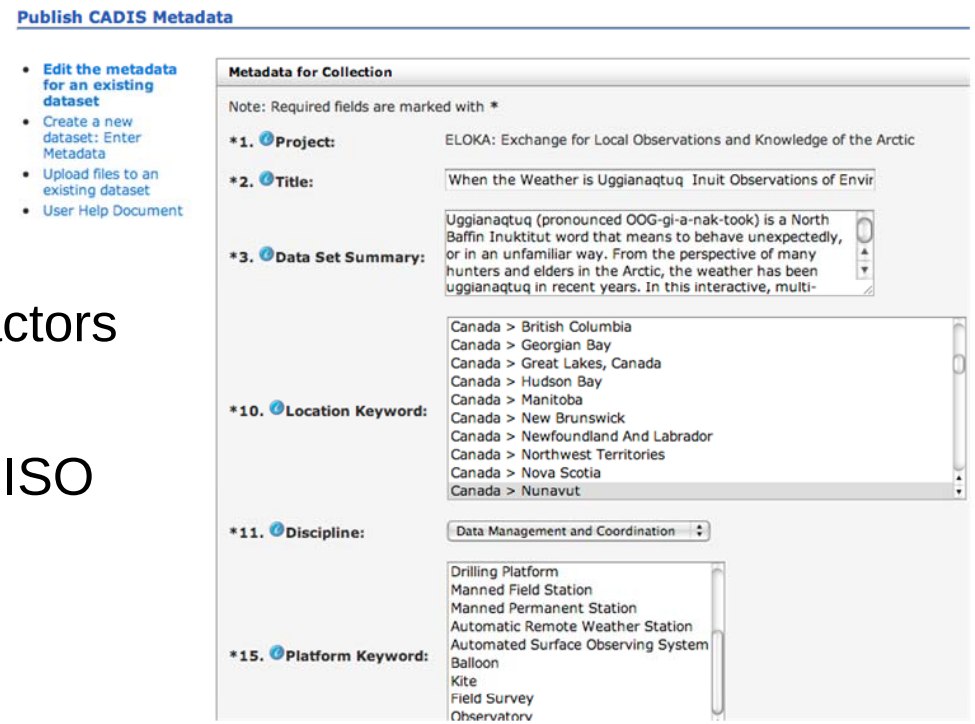
- Need to respond quickly to AON project needs
- Establish and interview a representative user group
- Developed CADIS Metadata Profile and use cases
- Leverage existing SGF data management tools
- Provide guidance and support for project investigators
- Develop data conversion tools and documented processes

CADIS Data Portal

SGF Gateway

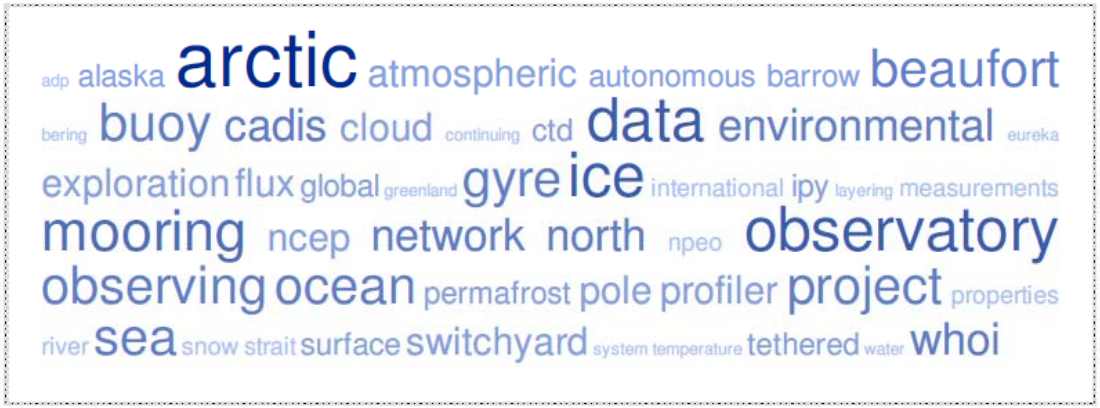
- Stand Alone Gateway Node
- Data Publication Web Interface
- Access Control and User Management
- Faceted and GeoSpatial Search, Collection Browsing
- OAI Metadata Exchange OAI-DIF (GCMD, Damocles)
- Data Access and Download (HTTP, wget)
- Metrics capture and funding agency reporting
- SGF Version 1.3 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, FCDG, ISO
- Profile driven UI with validation
- NASA GCMD vocabulary
- Curation by metadata team
- Integrated end user help documentation

CADIS 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

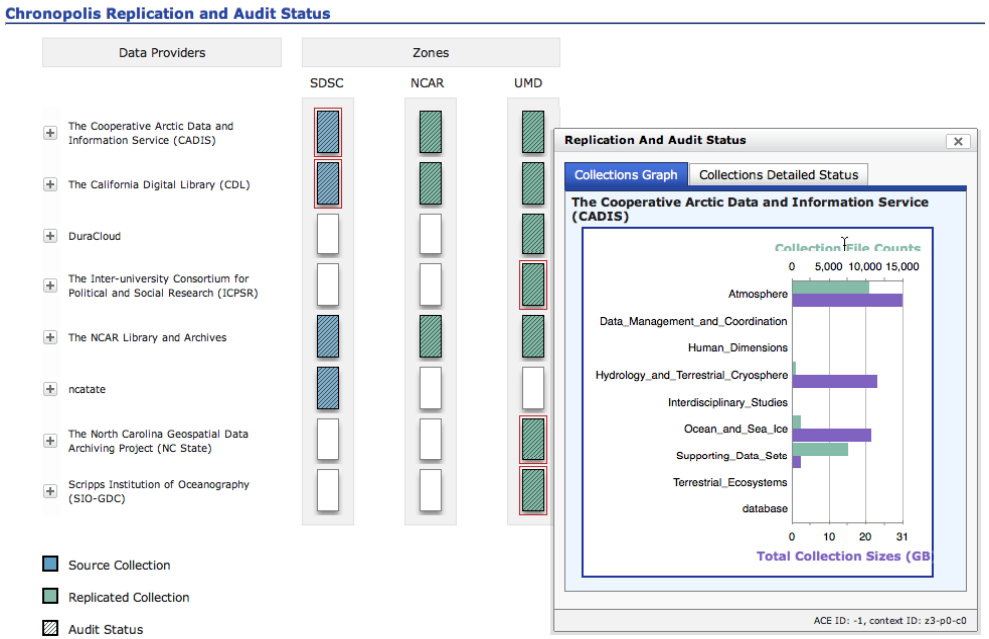
Integration Effort of Available Components

- 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?

Digital Preservation

Ensures Long Term Value

- At Risk Data Collections
- <Add More Detail>

CADIS Now – System Use

Content and use summary

----

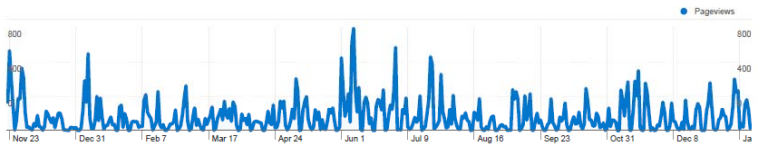
54 Projects
39 Lead Principle Investigators

----

355 Archived data collections
110 Brokered data collections
84Gb Archived files
2100 Unique Annual Visitors

aoncadis.ucar.edu
Content by Title

Nov 22, 2009 - Jan 21, 2011
Comparing to: Site



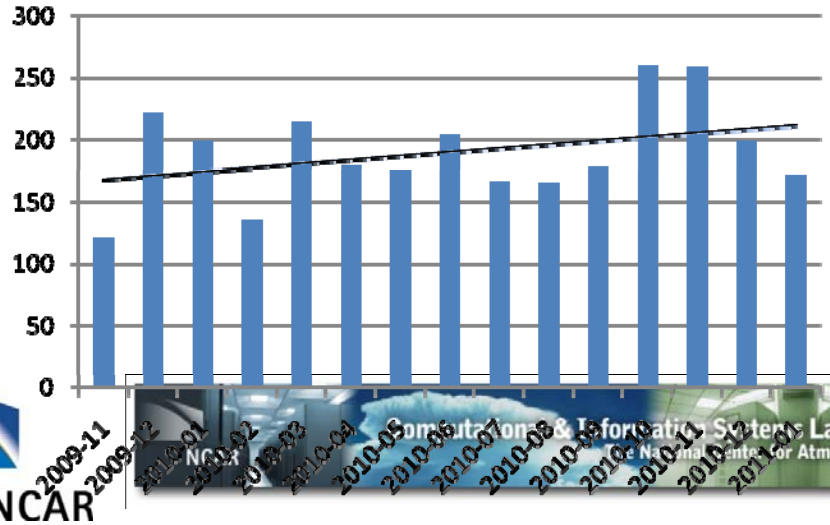
601 page titles were viewed a total of 46,613 times

Content Performance

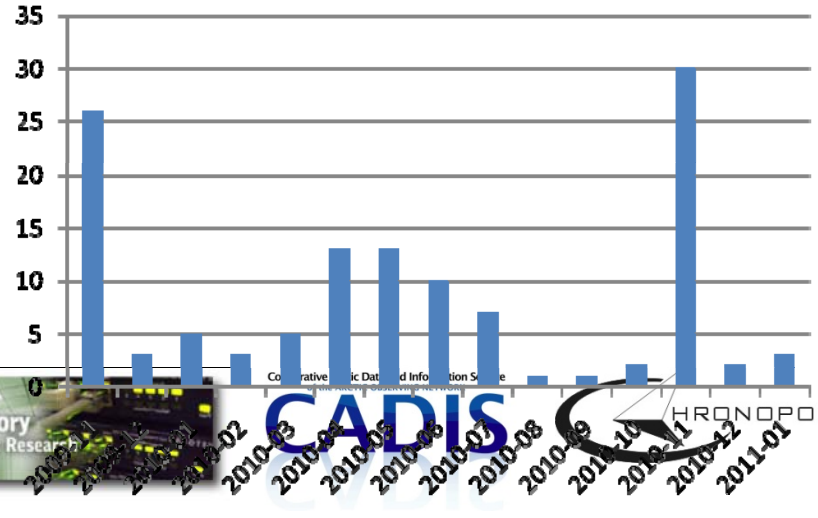
Pageviews	Unique Pageviews	Avg. Time on Page	Bounce Rate	% Exit	$ Index
46,613 % of Site Total: 100.00%	27,226 % of Site Total: 100.00%	00:01:07 Site Avg: 00:01:07 (0.00%)	41.43% Site Avg: 41.43% (0.00%)	11.49% Site Avg: 11.49% (0.00%)	$0.00 Site Avg: $0.00 (0.00%)

Page Title	Pageviews	Unique Pageviews	Avg. Time on Page	Bounce Rate	% Exit	$ Index
CADIS Gateway	6,107	3,980	00:01:30	41.99%	34.80%	$0.00
CADIS Advanced Search	4,303	1,193	00:01:01	51.60%	9.95%	$0.00
Download Data	2,546	1,586	00:01:31	52.56%	16.85%	$0.00
Publish	2,103	1,121	00:01:34	21.05%	3.99%	$0.00

Users per month



New Collections per month



NCAR

Computational & Information Systems Laboratory
The National Center for Atmospheric Research

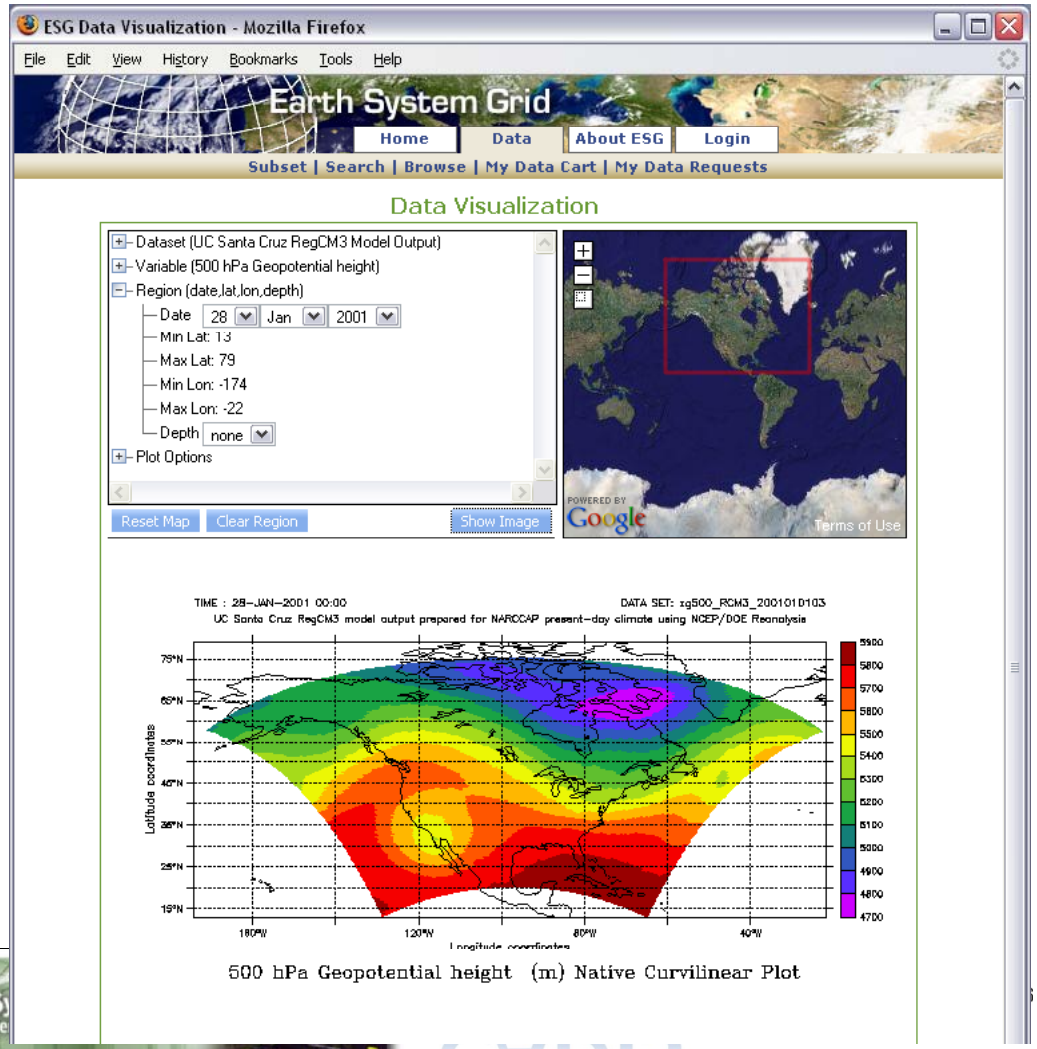
CADIS

HRONOPOLIS

User Interface for Visualization

Initial Prototype:

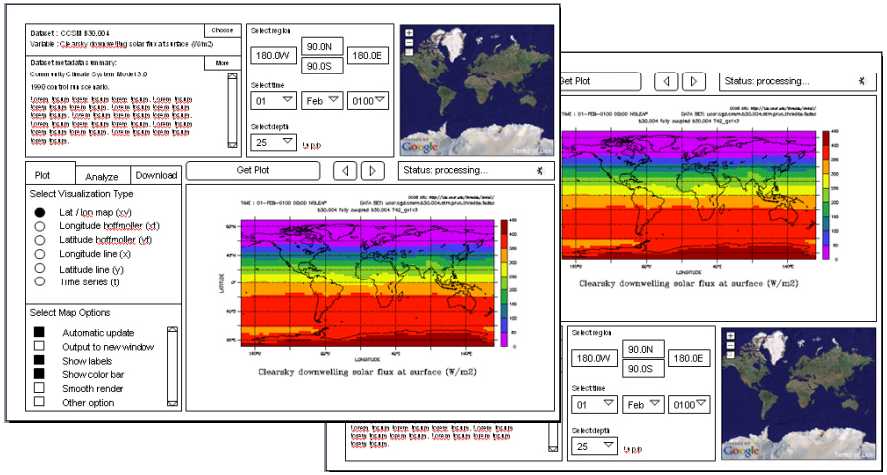
- Google Map API
- Custom selection overlay
- Dojo JavaScript Toolkit
- Tree based selection
- Ajax application with JSON
- Proxy client to access LAS
- LAS Armstrong v 0.4
- TDS v 3.16

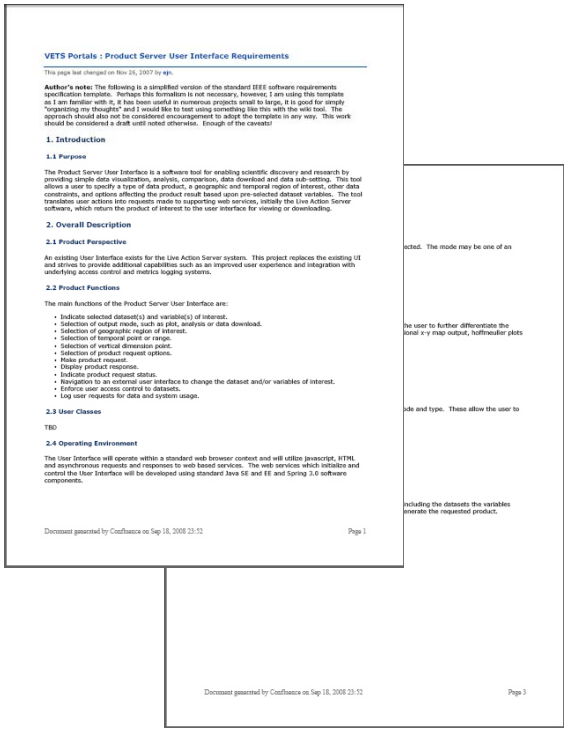


User Interface for Visualization

Back to the Drawing Board:

- Review of feedback
- Wire-frame drawings
- Basic formal specification





Next Steps:

What will we do with these findings?

- Collect more feedback from ESG test-bed deployment.
- Consider formalizing user studies and observations
- Expose more analysis features of LAS
- Add multi-variable selection for analysis, sub-setting, layering
- Integrate geographic selection with search tools
- Explore more “polar capable” UI components
- Explore WMS visualization capabilities
- Exploit more of the F-TDS analysis, sub-setting capabilities
- Continue to refine and improve quality of the UI
- Encourage community convergence on UI best practices

CADIS System Use

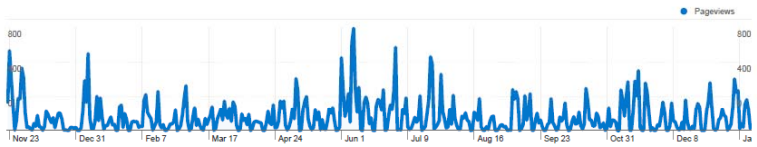
Content summary

64 AON Projects
Over 89 Lead and Collaborating Investigators

355 Archived data collections
110 Brokered data collections

aoncadis.ucar.edu
Content by Title

Nov 22, 2009 - Jan 21, 2011
Comparing to: Site

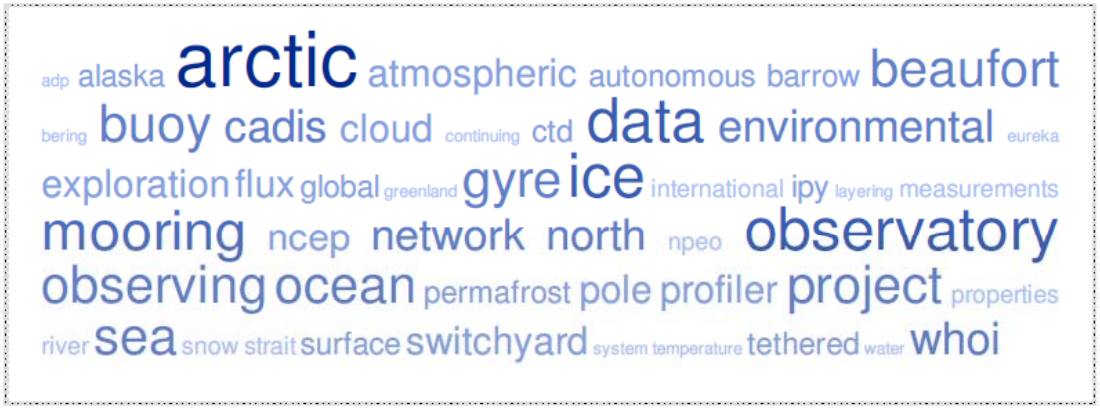


601 page titles were viewed a total of 46,613 times

Content Performance

Pageviews	Unique Pageviews	Avg. Time on Page	Bounce Rate	% Exit	$ Index
46,613 % of Site Total: 100.00%	27,226 % of Site Total: 100.00%	00:01:07 Site Avg: 00:01:07 (0.00%)	41.43% Site Avg: 41.43% (0.00%)	11.49% Site Avg: 11.49% (0.00%)	$0.00 Site Avg: $0.00 (0.00%)

Page Title	Pageviews	Unique Pageviews	Avg. Time on Page	Bounce Rate	% Exit	$ Index
CADIS Gateway	6,107	3,980	00:01:30	41.99%	34.80%	$0.00
CADIS Advanced Search	4,303	1,193	00:01:01	51.60%	9.95%	$0.00
Download Data	2,546	1,586	00:01:31	52.56%	16.85%	$0.00
Publish	2,103	1,121	00:01:34	21.05%	3.99%	$0.00



NCAR

Computational & Information Systems Laboratory
The National Center for Atmospheric Research

Cooperative Arctic Data and Information Service of the ARCTIC OBSERVING NETWORK
CADIS

CHRONOPOLIS