ACADIS Advanced Cooperative Arctic Data & Information Service

ACADIS, Analytic Potential, and Stewardship Maturity

Ruth Duerr, Toni Rosati

Outline

- Background
- Definitions
- ACADIS
  - Stewardship maturity matrix
  - Analytic potential

Background

- Peng's Presentation at 2014 summer ESIP meeting
- Data Stewardship committee agreed on assess general utility
- ACADIS assessment

ACADIS

Advanced Cooperative Arctic Data & Information Service

NSF

NSIDC
National Snow and Ice Data Center
University of Colorado Boulder

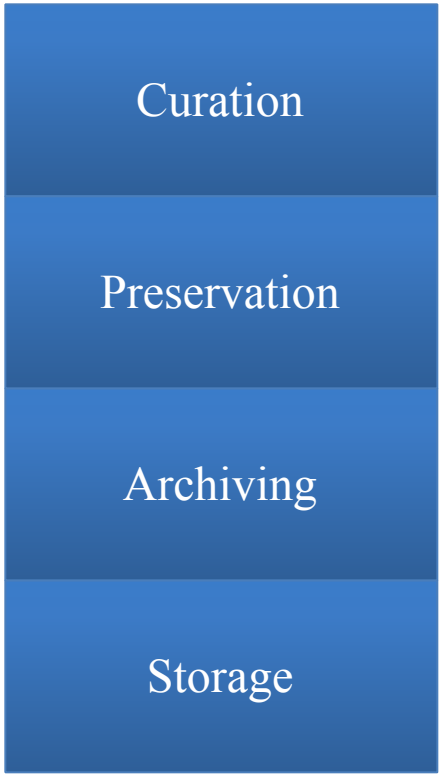
UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Definitions

Term	Definition	Implications
Curation	***Processes that add value to foster discovery and reuse***	***Allows others to reuse data; supports cross-disciplinary research***
Preservation	***Providing enough representation information, context, metadata, fixity, etc. to support use and interpretation by agents other than the original data producer***	***Ability to use own data in the future (e.g., 5 yrs out)***
Archiving	***Activities that enable long-term retention of digital materials (e.g., replication, fixity, identifiers, etc.)***	***Provides identifiers for sharing, citation, etc.***
Storage	***Basic physical storage with backup and restore services***	***Allows basic sharing***

Curation Stack Model



Choudhury, G. S., Palmer, C. L., Baker, K. S., & DiLauro, T. (2013, January). Levels of services and curation for high-functioning data. Presented at the International Digital Curation Conference, Amsterdam, Netherlands.

ACADIS Data Stewardship Research Study

Purpose: Bringing together ACADIS metadata, data, and preservation work in the context of future scientific use

Methods (part 1):

- Apply NOAA Stewardship Maturity Matrix (Peng et al, 2015) to ACADIS catalog entries
  - Goal is to provide a "systematic framework to assess the vigor of stewardship practices applied to individual environmental datasets [and] to provide consistent information on data quality, data integrity and usability to users and stakeholders"
  - Does not in and of itself provide information on the potential utility of the data
- Assess analytic potential (Palmer et al, 2011) of these data sets
  - Assessment of:
    - Number of potential user communities (including policy-makers, general public)
    - Fitness for purpose
    - Preservation readiness
  - Special circumstances also noted (e.g., long-time series / aggregation potential)

Application to ACADIS data: General

Methods (part 2):

- Spreadsheet of all data sets in the ACADIS catalog obtained (2851 data sets in total)
  - 37 have no data in the ACADIS system and no links indicating where the data might be (only metadata are available)
  - 152 have no data in the ACADIS system (includes the 37 above)
  - 814 have data in the ACADIS gateway
  - 1,885 have data in EOL
- Systematic sampling used (every 11th sample to be examined in detail)
  - Review gateway metadata record (completeness, descriptive value)
  - If data link available, follow it (data format, data documentation, services available, etc.)
  - If Readme or other metadata or documentation is available, read it (is the content of the data understandable, is the quality described, etc.)

Application to ACADIS data: NOAA Maturity Matrix

Methods (part 2):

- Initial application of the Matrix
  - It should be noted that the paper was not published until mid-January, 2015
  - NOAA intends to publish a template for assessing a data set; but this has not yet been done
  - Many columns in the matrix depend on specific sections of metadata records that ACADIS does not have or are based on assumptions which are not appropriate for most investigator led research projects
    - These fields were reinterpreted to be meaningful for investigator led science projects
    - Feedback on these results will be taken back to the ESIP Data Stewardship Committee

Very Preliminary Results with the Maturity Matrix (1 of 3)

- A mere 37 data sets have been assessed so far (~1.3%) so findings are not generally statistically meaningful yet
- Preservability
  - Defined as the degree of conformance to OAIS Reference Model and NARA archiving standards
  - Set to 0 for all data not actually held by ACADIS systems
  - Set to 3 (i.e., archived in a designated repository/archive with redundancy, and conforming to community archiving process and metadata) for all data held in ACADIS systems
- Accessibility
  - Based on type and granularity of search supported and whether data is available on-line
  - Set to 3 for all ACADIS data, since collection/data set level search is supported through the ACADIS Gateway but granule/file level search is not

Very Preliminary Results with the Maturity Matrix (2 of 3)

Usability

- Based on knowledge required to understand the data, the metadata standard and content, the data format and content, and what access services (e.g., subsetting) are provided
- Separated into 4 fields measured separately (1-5 scale) and then averaged:
  - Knowledge required to understand data format and content: 0 used if data not available through ACADIS
    - Full range of values observed
  - Metadata
    - Maximum value of 3 so far
  - Data documentation
    - Full range of values observed
  - Access services
    - Maximum value of 2 so far

Data Quality Assurance

- Measures taken to ensure data meets specifications
- Not particularly relevant for many measurements/observations
- Restricted definition to include actions by repository to verify provided data (format, viruses, manifests, checksums, file readability, etc.) on ingest
- Set to 1 for all ACADIS data

Very Preliminary Results with the Maturity Matrix (3 of 3)

Data Quality Control and Monitoring

- Measures taken to determine whether data has issues (gaps, outliers, etc.) and handle any issues detected
- Not particularly relevant for many measurements/observations
- Restricted definition to include actions by investigator to record and document QC procedures, fill gaps, etc.
- Full range of values observed

Application to ACADIS data: Analytic Potential

- No rubric or measurement scale exists for the 3 axes here
- Researching various measurement criteria is required
- Current criteria:
  - User communities
    - Number of potential research communities
    - Do decision makers care about these data (assuming they were cast in an appropriate form)?
    - Does the public care?
  - Fitness for purpose
  - Preservation readiness
- Very early results
  - Number of potential research communities can almost always be justified to be large – so this is not that useful a discriminator in its current form
  - Fitness for purpose is a function of the community

Other Observations

- Some data sets held externally have comprehensive metadata in a non-ACADIS compatible format (e.g., EML, DDI, etc.)
- Not all data sets held by ACADIS actually contain data (e.g., pdf or gif images of data, reports of various types, etc.) – a usability issue; it would be useful to have mechanisms to screen these out as needed
- Some ACADIS data sets are also held elsewhere (not necessarily a bad thing)
- Some data sets are duplicated in the system due to multiple file formats being treated as different data sets rather than different representations of the same data – a discoverability and assessment issue for users

References

- Peng, Ge, Jeffrey L. Privette, Edward J. Kearns, Nancy A. Ritchey, and Steve Ansari. 2015. “A Unified Framework for Measuring Stewardship Practices Applied to Digital Environmental Datasets.” *Data Science Journal* 13: 231–53. doi:10.2481/dsj.14-049.
- Palmer, Carole L., Nicholas M. Weber, and Melissa H. Cragin. "The analytic potential of scientific data: Understanding re−use value." *Proceedings of the American Society for Information Science and Technology* 48, no. 1 (2011): 1-10.

Questions?