

## **National Snow and Ice Data Center Distributed Active Archive Center**

### ***Data Deaccessioning Guidelines for the NSIDC DAAC***

**Sheena Beaverson**

**June 29, 2009**

#### **Introduction**

I have been working to complete a Practicum at the National Snow and Ice Data Center (NSIDC) in Boulder, Colorado under the guidance of data stewardship managers of the NSIDC Distributed Active Archive Center (DAAC). "The NSIDC DAAC archives and distributes brightness temperature data, polar atmosphere data, satellite imagery, sea ice data, snow cover data, and ice sheet data. Data [have been] derived from a range of operational and historical sources that use widely disparate technologies" (see <http://nsidc.org/daac/>). In addition, "[m]any data sets come to NSIDC as a result of PI offerings that are unrelated to any program" (NSIDC, 2005 pg. 4). The focus of my Practicum Special Project has been to develop a draft guideline for data de-accessioning, to be used at NSIDC DAAC.

A logical approach to developing a data deaccessioning tool for evaluation of NSIDC DAAC data resources was to adapt the recently completed *Data Acceptance Plan for the NSIDC DAAC* (Weaver and Duerr, 2009) with modifications taken from current best practices at NSIDC and other national or international agencies. Descriptive information about data resources that have not been previously assessed in a formal manner must be recorded, prior to disposal of data resources.

#### **Supplement to Data Acceptance Plan for the NSIDC DAAC**

In a 2005 internal document entitled *NSIDC Procedures for Long-Term Data Management* the authors proposed that "[t]he possibility of future data retirement or deletion [should be] covered with the contributing agency or individual at the time of data donation" (NSIDC, 2005 pg. 6). Nevertheless, the NSIDC DAAC guideline for submitting data does not directly address the potential for future deaccessioning of accepted data resources (NSIDC DAAC, 2009). I propose the addition of a limited number of questions to the Submission Template that pertain this topic. These would include:

- What was the cost of data acquisition?
- What would be the projected cost to recollect or recreate the data?  
Would it actually be possible to recollect or recreate the data?
- What are the assumptions about the level of service for these data in the future?  
Will these data be relevant 5 years from now? 10 years? 50 years?
- What triggering event would cause these data to be considered for disposal?  
(i.e. scheduled review, staff retirement, software or hardware obsolescence, etc.)
- Are these data also being maintained by another agency? If so, provide details.

## **Data Deaccessioning Process Steps**

The following process steps have been developed from the section entitled Data Acceptance Process on pages 7 through 10 of the *Data Acceptance Plan for the NSIDC DAAC* (Weaver and Duerr, 2009) and also the NOAA Process Steps outlined on page 6 of the *NOAA Procedure for Scientific Records Appraisal and Archive Approval Guide for Data Managers* (National Oceanic and Atmospheric Administration, 2008).

Consideration of data for deaccessioning would trigger an in-depth review of all aspects of a given data set and its related resources. Documentation and analysis of the data, related resources, and deaccessioning appraisal and evaluation is guided by the *Data Deaccessioning Assessment Guideline for the NSIDC DAAC*, presented in Appendix A of this document. Other agencies wishing to utilize this guideline are directed to the companion *Data Deaccessioning Assessment Guideline for any Agency* contained in Appendix B. Appendix C includes a listing of references from which conceptual or actual guidelines were obtained during the development of appendices A and B.

It is imperative that technical data managers and scientific staff each be included in conducting a data deaccessioning assessment. It is assumed that a group of individuals charged with evaluating sophisticated data resources for disposal would be similar in composition to NSIDC Product Teams.

“Most of the data activities in the DAAC are handled through Product Teams, which typically include representatives from each of the specialty areas at NSIDC. In general a product team will include a technical writer, a user services representative, a scientist, a scientific programmer, a data operations representative, a product team lead, and other technical personnel as needed” (Weaver and Duerr, 2009 pg. 8).

### **Step 1: Triggering Event for Consideration for Deaccessioning**

Data resources that have been extant in the agency for a long period time may be subjected to a triggering event which will cause them to be evaluated for disposal. Possible triggers might include, but are not restricted to, the following: a predetermined date for resource review, retirement or advancement of key staff members, impending obsolescence of system hardware or software components, conclusion of a contractually required period of support at the current level of service, or termination of funding for data maintenance and access.

A triggering event would cause the assignment of the appropriate Product Team to the deaccessioning appraisal process.

### **Step 2: Gather Existing Documentation (Appendix A, Section 1)**

It is appropriate for the Product Team leader to gather and disseminate to the Product Team any previous evaluations and existing documentation that pertains to the content and management of the data collection being assessed. Potential information sources are outlined in the Section 1 of the guideline in Appendix A. Previous data appraisal records, metadata, web resources, manuals and how-to guides, and reports detailing use metrics are all potential information sources.

**Step 3: Information Retrieval**  
**(Appendix A, Sections 3 through 8)**

Once all existing documentation is compiled, the Product Team should meet to determine what information is contained in the documentation and what information will need to be ascertained. The team will then assign members to track down outstanding facts in response to the topical questions outlined in Sections 2 through 8 of the *Data Deaccessioning Assessment Guideline* in Appendix A. Section 2 of this guideline is, in fact, a duplicate copy of questions outlined in the *Data Acceptance Plan for the NSIDC DAAC; Appendix C, Section 1*. These are the first items which need to be documented for a data resource, assuming that the data were not previously accessioned by NSIDC utilizing this guideline. Sections 3 through 8 address the following:

- Section 3: Uniqueness
- Section 4: Data File Format, Hardware and Software System Components
- Section 5: Data Content and Nature of the Data
- Section 6: Costs and Financial Considerations
- Section 7: Obligations
- Section 8: Opportunities

**Step 4: Ranking and Recommendation**  
**(Appendix A, Sections 9 and 10)**

Once the data resource is fully documented, the Product Team will meet to establish a ranking, Section 9, of the scientific relevance of the data contents and the requirements for on-going resource support. This will, in turn, guide the Product Team to make a recommendation (Section 10). Team members will either recommend that work be done to establish a new level of service, that the data is maintained at the current level of service, or that the data is deaccessioned from NSIDC DAAC.

**Step 5: Administrative Review and Execution of Recommendation**  
**(Appendix A, Section 11)**

The completed *Data Deaccessioning Assessment Guideline*, complete with dataset ranking and ultimate recommendation, should then be submitted to the Polar DAAC Advisory Group (PODAG) for administrative review.

Once clearance is obtained from PODAG, an archive of all documentation of this assessment, the decision, and PODAG Approval should be established. This is especially necessary to justify a recommendation for permanent data disposal in the future.

The recommended actions for the data resources should then be implemented.

## References

National Snow and Ice Data Center, 2005. NSIDC Procedures for Long-Term Data Management. National Snow and Ice Data Center. (*internal document*).

Weaver, R., and Duerr, R., 2009. Data Acceptance Plan for the NSIDC DAAC. (*internal document*).

## **Appendix A: Data Deaccessioning Assessment Guideline for the NSIDC DAAC**

### ***Section 1: Previous Evaluations and Existing Documentation***

- Why have these data been targeted for disposal? Identify the triggering event.
  - (i.e. scheduled review, staff retirement, software or hardware obsolescence, etc.)
- Is there documentation about the initial decision to archive and maintain?
- Have the data been formally appraised?
  - With *Data Acceptance Plan for the NSIDC DAAC Submission Template*?
  - With the *USGS EROS Records Appraisal Tool*?
  - With some other appraisal tool?
- Is there documentation that describes the nature of the data? List these files.
  - (i.e. formal metadata, web resources, manuals, how-to guides)
  - (provide server/directory/filename path or appropriate web address)
- Which staff member(s) are currently assigned to manage the data?
- Which Product Team will be assigned to evaluate data for potential disposal?
  - (i.e. staff member(s) and scientist(s))
- What is the start date of the data evaluation process?

### ***Section 2: Data Acceptance Plan for the NSIDC DAAC; Appendix C, Section 1***

*(to be filled out if NSIDC DAAC data not previously evaluated in a formal manner)*

- Title, relevant acronym(s) or abbreviations, and Earth Science Data Type (ESDT)
- PI name, organization and contact information:
- NSIDC DAAC Tracking Number:
- Brief Description or Abstract
- Product algorithm theoretical basis
- Publication List
  - (list of published works related to the data set)
- PI Justification
- Data Acquisition Methods
- Quality and Accuracy Information
  - (cal/val, relative and absolute uncertainty, stability, maturity of algorithm)
- Product Use
  - (include any limitations on use)
- Ingest Frequency and Scope
  - 1-time Ingest? Yearly Ingest? Ongoing Ingest?
  - Number of files per ingest session

### ***Section 3: Uniqueness***

- Are the data also maintained by the originating, or any other, agency?
- Is this a unique or a duplicate copy of the data?

#### ***Section 4: Data File Format, Hardware and Software System Components***

- Data File Format
  - How many files make up data collection?
  - Is there a formal file naming convention applied?
  - What is the average file size and total volume of data?
  - Are the files still readable?
  - Are the data files uncompressed or compressed?
  - Is the file format stable, appropriate, etc.
  - Describe the digital data storage medium. (CDROM, DVD, hard drive...)
  - Describe the data processing requirements.
- Are the data solely in a digital format, or is there also an analog version?
- Describe the current IT system hardware involved in maintaining the data.
- List all required software and customized programs related to utilizing the data.
- Describe the IT implications for long-term data maintenance.
  - (i.e. file format migration, IT system upgrades, access, user support, etc.)

#### ***Section 5: Data Content and Nature of the Data***

- Is this a primary data resource (raw, or unprocessed, data)?
- Is this a derivative data resource (PR item/ past version of data/ subset of data)?
- Detail the heritage, or provenance, of the data; provide any related justifications.
  - Where did this data resource come from (agencies or individuals)?
  - Who authorized NSIDC to manage these data (EOS, DAAC, etc.)?
  - Is there a DAAC Justification?
  - Are there User Working Group, MOUs, requests, or other justifications?
  - Are the data relevant to NSIDC DAAC Ongoing Efforts?
- Estimate the completeness of the data and any relevant supplemental resources.
- Evaluate the integrity of the data resource.
  - Are the data authentic, reliable, unaltered and usable?
- What is the current scientific relevance or utility of the data?
  - (i.e. a Climate Data Record or has Earth Science Data Record Potential)
- What is the projected scientific relevance or utility of the data resources?
  - What is the likelihood that the data will be in high demand in the future?
- Are the data supplemental to other scientific resources or research?
- Describe the spatial area and temporal ranges covered by the data.
- Does this data have special intrinsic relevance that should be noted?
  - (i.e. increase the value of other data, artistic merit, historically significant)

#### ***Section 6: Costs and Financial Considerations***

- What was the original funding source and cost of data acquisition?
- What was the original funding source and amount put toward data management?
- What would be the projected cost to recollect or recreate the data?
  - Would it actually be possible to recollect or recreate the data?
- Identify the current NSIDC DAAC Level of Service. Is this an appropriate level?
- Estimate current annual costs to maintain, archive, process and support users.
- Estimate the cost to generate documentation, if no documentation exists.
- Estimate the cost to make the data accessible, if not already accessible.

### **Section 7: Obligations**

- Are their obligations for on-going support of these data?
  - (i.e. legal requirements, contract-driven obligations, sharing agreements)
- Is the data in alignment with NSIDC DAAC policies?
  - (i.e. agency mission, collection scope, and data management policies)
- Were there informal obligations, agreed to or implied, at the time of acquisition?
  - (i.e. an outlook of the data as being permanent or temporary)
- Are the data of a proprietary, sensitive, or classified nature?
- Would the data be of use in defending this agency against charges of data fraud?

### **Section 8: Opportunities**

- Do the data serve to earn prestige or expand the NSIDC DAAC user base?
  - Is there measurable activity (use metrics) for accessible resources?
  - Were the data frequently referenced in journal articles?
  - Were the data featured in news articles or blogs?
- Do the data still serve the original purpose or new purposes?
- Do the data attract or strengthen bonds with funding providers?
- Do the data encourage engagement with NSIDC DAAC by other agencies?

### **Section 9: Ranking**

Category	Desired State	Work Difficulty
Uniqueness	Desired State:	
	Additional Factors:	
	Total	
Technical Implications	Desired State:	
	Additional Factors:	
	Total	
Data Content and Scientific Importance	Desired State:	
	Additional Factors:	
	Total	
Cost Implications	Desired State:	
	Additional Factors:	
	Total	
Obligations	Total	
Opportunities	Desired State:	
	Additional Factors:	
	Total	
	Grand Total	

### ***Section 10: Recommendations***

- Is it recommended that work be done to establish a new level of service?
  - What data specialist would be assigned this work?
- Is it recommended that the data be maintained at the current level of service?
  - Should a future action or reassessment be scheduled?
- Is it recommended that the data be deaccessioned from NSIDC DAAC?
  - Would a decision to purge need to be open for public comment or review?
  - Should other repositories be contacted to determine interest in a transfer?
    - (i.e. USGS EROS Purge Alert, interagency contacts, etc.)
  - Can permanent disposal be completed in compliance with NSIDC procedures for managing digital content?

### ***Section 11: Final Actions***

- Submit to PODAG for approval or rejection of recommendation.
- Archive documentation of this assessment, the decision, and PODAG Approval
- Document the assessment completion date and recommended action(s).
- Implement recommended action(s).



## **Appendix B: Data Deaccessioning Assessment Guideline for any Agency**

### ***Section 1: Previous Evaluations and Existing Documentation***

- Why have these data been targeted for disposal? Identify the triggering event.
  - (i.e. scheduled review, staff retirement, software/ hardware obsolete, etc.)
- Is there documentation about the initial decision to archive and maintain?
- Have the data been formally appraised?
  - With *Data Acceptance Plan for the NSIDC DAAC Submission Template*?
  - With the *USGS EROS Records Appraisal Tool*?
  - With some other appraisal tool?
- Is there documentation that describes the nature of the data? List these files.
  - (i.e. formal metadata, web resources, manuals, how-to guides)
  - (provide server/directory/filename path or appropriate web address)
- Which staff member(s) are currently assigned to manage the data?
- Which staff member(s) will be assigned to evaluate data for potential disposal?
- Which scientist(s) will be assigned to evaluate data for potential disposal?
- What is the start date of the data evaluation process?

### ***Section 2: General Information***

- Title, relevant acronym(s) or abbreviations, tracking number, etc.
- PI name, organization and contact information.
- Brief description or abstract.
- Publication list; list published works related to the data set.
- Product use statement, (include any limitations on use).
- Ingest Frequency (1-time? Yearly? Ongoing?) and number of files per ingest.

### ***Section 3: Uniqueness***

- Are the data also maintained by the originating, or any other, agency?
- Is this a unique or a duplicate copy of the data?

### ***Section 4: Data File Format, Hardware and Software System Components***

- Data File Format
  - How many files make up data collection?
  - Is there a formal file naming convention?
  - What is the average file size and total volume of data?
  - Are the files still readable?
  - Are the files uncompressed or compressed?
  - Is the file format stable, appropriate, etc.
  - Describe the digital data storage medium. (CDROM, DVD, hard drive...)
  - Describe the data processing requirements.
- Are the data solely in a digital format, or is there also an analog version?
- Describe the current IT system hardware involved in maintaining the data.
- List all required software and customized programs related to utilizing the data.
- Describe the IT implications for long-term data maintenance.
  - (i.e. file format migration, IT system upgrades, access, user support, etc.)

### ***Section 5: Data Content and Nature of the Data***

- Is this a primary data resource (raw, or unprocessed, data)?
- Is this a derivative data resource (PR item/ past version of data/ subset of data)?
- Detail the heritage, or provenance, of the data; provide any related justifications.
  - Where did this data resource come from (agencies or individuals)?
  - Who authorized our agency to manage these data?
  - Are the data relevant to our agency's ongoing efforts?
- Estimate the completeness of the data and any relevant supplemental resources.
- Evaluate the integrity of the data resource.
  - Are the data authentic, reliable, unaltered and usable?
- What is the current, and projected, scientific relevance or utility of the data?
  - What is the likelihood that the data will be in high demand in the future?
- Are the data supplemental to other scientific resources or research?
- Describe the spatial area and temporal ranges covered by the data.
- Does this data have special intrinsic relevance that should be noted?
  - (i.e. increase the value of other data, artistic merit, historically significant)

### ***Section 6: Costs and Financial Considerations***

- What was the original funding source and cost of data acquisition?
- What was the original funding source and amount put toward data management?
- What would be the projected cost to recollect or recreate the data?
  - Would it actually be possible to recollect or recreate the data?
- Identify the current level of service. Is this an appropriate level?
- Estimate current annual costs to maintain, archive, process and support users.
- Estimate the cost to generate documentation, if no documentation exists.
- Estimate the cost to make the data accessible, if not already accessible.

### ***Section 7: Obligations***

- Are their obligations for on-going support of these data?
  - (i.e. legal requirements, contract-driven obligations, sharing agreements)
- Is the data in alignment with our agency's policies?
  - (i.e. agency mission, collection scope, and data management policies)
- Were there informal obligations, agreed to or implied, at the time of acquisition?
  - (i.e. an outlook of the data as being permanent or temporary)
- Are the data of a proprietary, sensitive, or classified nature?
- Would the data be of use in defending this agency against charges of data fraud?

### ***Section 8: Opportunities***

- Do the data serve to earn prestige or expand our agency's user base?
  - Is there measurable activity (use metrics) for accessible resources?
  - Were the data frequently referenced in journal articles?
  - Were the data featured in news articles or blogs?
- Do the data still serve the original purpose or new purposes?
- Do the data attract or strengthen bonds with funding providers?
- Do the data encourage engagement with our agency by other agencies?

## Section 9: Ranking

Category	Desired State	Work Difficulty
Uniqueness	Desired State:	
	Additional Factors:	
	Total	
Technical Implications	Desired State:	
	Additional Factors:	
	Total	
Data Content and Scientific Importance	Desired State:	
	Additional Factors:	
	Total	
Cost Implications	Desired State:	
	Additional Factors:	
	Total	
Obligations	Total	
Opportunities	Desired State:	
	Additional Factors:	
	Total	
	Grand Total	

## Section 10: Recommendations

- Is it recommended that work be done to establish a new level of service?
  - What data specialist would be assigned this work?
- Is it recommended that the data be maintained at the current level of service?
  - Should a future action or reassessment be scheduled?
- Is it recommended that the data be deaccessioned from our agency?
  - Would a decision to purge need to be open for public comment or review?
  - Should other repositories be contacted to determine interest in a transfer?
    - (i.e. USGS EROS Purge Alert, interagency contacts, etc.)
  - Can permanent disposal be completed in compliance with our agency 's procedures for managing digital content?

## Section 11: Final Actions

- Submit to agency managers for approval or rejection of recommendation.
- Archive documentation of this assessment, the decision, and manager approval.
- Document the assessment completion date and recommended action(s).
- Implement recommended action(s).

## Appendix C: Reference Materials for Appendices A and B

Beagrie, Neil, Chruszcz, Julia and Lavoie, Brian, 2008. Keeping Research Data Safe, a Cost Model and Guidance for UK Universities. Joint Information Systems Committee.  
<http://www.jisc.ac.uk/publications/documents/keepingresearchdatasafe.aspx>

Broome, John and Moir, Phil, 2008. ESS Data Set Stewardship. ESS Data Management and Dissemination Branch. Version 8.0. (*internal document*).

Committee on Earth Observation Satellites (CEOS), 2009. Web Page for the Announcement of Datasets Being Purged by Agencies or Organizations. Last updated January 5, 2009; accessed June 15, 2009.  
[http://edc.usgs.gov/archive/ceos/data\\_purge\\_alert.html](http://edc.usgs.gov/archive/ceos/data_purge_alert.html)

Day, Michael, 2004. The Selection, Appraisal and Retention of Digital Scientific Data: the ERPANET / CODATA workshop. Ariadne. Issue 39. April 30, 2004.  
<http://www.ariadne.ac.uk/issue39/erpanet-rpt/>

Digital Curation Centre, 2009. DCC Data Curation Lifecycle Model. Last updated June 4, 2009, accessed June 15, 2009.  
<http://www.dcc.ac.uk/lifecycle-model/>

Faundeen, John, 2009. Regarding the EROS Data Management Policy. June 16, 2009. (*personal communication*).

Faundeen, John, 2008. EROS Records Management Plan. Version 3.1.  
[http://edc.usgs.gov/government/records/media/EROS\\_Records\\_Management\\_Plan.doc](http://edc.usgs.gov/government/records/media/EROS_Records_Management_Plan.doc)

Faundeen, John L., 2003. *The Challenge of Archiving and Preserving Remotely Sensed Data*. Data Science Journal, Volume 2, 31 October 31, 2003.  
[http://www.jstage.jst.go.jp/article/dsj/2/0/2\\_159/\\_article](http://www.jstage.jst.go.jp/article/dsj/2/0/2_159/_article)

Library and Archives Canada, 2007. Canadian Digital Information Strategy. Draft Version of October, 2007.  
<http://www.collectionscanada.gc.ca/obj/012033/f2/012033-1000-e.pdf>

National Archives of Scotland, 2009. Information Management Policy for the Digital Data Archive and Future EDRMS. Last updated January 27, 2009; accessed June 16, 2009.  
<http://www.nas.gov.uk/recordKeeping/informationManagementPolicy.asp>

National Oceanic and Atmospheric Administration, 2008. NOAA Procedure for Scientific Records Appraisal and Archive Approval Guide for Data Managers. August 15, 2008.  
[http://www.nosc.noaa.gov/docs/products/NOAA\\_Procedure\\_document\\_final\\_12-16-1.pdf](http://www.nosc.noaa.gov/docs/products/NOAA_Procedure_document_final_12-16-1.pdf)

National Snow and Ice Data Center, 2005. NSIDC Procedures for Long-Term Data Management. National Snow and Ice Data Center. (*internal document*).

Society of American Archivists, 2009. Winter 2009 Newsletter. Last updated December 19, 2008; accessed June 15, 2009.

<http://www.archivists.org/saagroups/acq-app/newsletter.asp#Deaccessioning>

U.S. Department of Housing and Urban Development, 2002. Data Requirements Document Checklist. April 10, 2002.

<http://nhl.gov/offices/cio/sdm/devlife/tempchecks/drchecklist.doc>

U.S. Geological Survey (USGS) National Center for Earth Resources Observation and Science (EROS), 2007. Records Appraisal Tool: Appraisal Questions. Last updated January 14, 2007; accessed June 22, 2009.

[http://eros.usgs.gov/government/ratool/view\\_questions.php](http://eros.usgs.gov/government/ratool/view_questions.php)

U.S. Geological Survey (USGS) National Center for Earth Resources Observation and Science (EROS), 2005. Long-Term Archive Acceptance Checklist. May 1, 2005.

<http://eros.usgs.gov/government/records/USGS%20LTA%20Requirements%20Checklist.pdf>

Van Wingen, R.S., Hathorn, F., and Sprehe, J. T., 1999. Principles for Information Technology Investment in U.S. Federal Electronic Records Management. *Journal of Government Information*, v26 n1 p33-42 Jan-Feb 1999.

Weaver, R., and Duerr, R., 2009. Data Acceptance Plan for the NSIDC DAAC. (*internal document*).