DATA ACCEPTANCE PROCEDURES AND LEVELS OF SERVICE AT THE NATIONAL SNOW AND ICE DATA CENTER

R. E. Duerr, R. L. Weaver

National Snow and Ice Data Center
Cooperative Institute for Research in Environmental Science
University of Colorado at Boulder

1. INTRODUCTION

All data are not created equal. Individual data sets vary from each other in a multitude of ways - from easily measured ways such as size, format, and complexity; to ways that require a more nuanced understanding, such as in the breadth and depth of a data set's potential user base, its "designated community" to use the terminology of the Reference Model for an Open Archival Information System [1]. Given limitations in the resources available, it is not surprising then that repositories, such as those of the National Snow and Ice Data Center, need to make choices about the level of support or services provided for each data set acquired. Such choices prioritize center activities. While such decisions have been an implicit part of NSIDC activities for many years, an effort was recently taken to explicitly define the Levels of Service supported at NSIDC. Baseline Levels of Service have been defined for all existing NSIDC data sets. In addition, Levels of Service considerations are a major component of the NSIDC Distributed Active Archive Center's (DAAC's) new data accessioning processes. In this paper we describe the Levels of Service currently supported at NSIDC and factors that affect the effort required to obtain a given level of service. We also discuss the process users should use if they wish to request that the NSIDC DAAC archive their data.

2. BACKGROUND

The basis for most research papers published in the Earth sciences today is data - be it data acquired by an investigator as part of their research perhaps as a result of a field campaign, data obtained from one of the many remote sensing systems or observing networks operated by a governmental agency or other organization, or data obtained through access to one of the many models currently in use. However, the days where it was possible to include all the data needed in order to replicate the results of that research in the peer-reviewed literature are long gone, if indeed they ever existed. Yet, results replication is a core tenet of the scientific process - at the heart of the reason for public trust in and consequently funding for scientific research. As a result of increasing recognition of this disconnect, the global scientific community is beginning to call for submission and archival of

research data to recognized archives (e.g., [2], [3]).

As perhaps a logical consequence of this, archives and data centers, such as the National Snow and Ice Data Center (NSIDC), are increasingly being inundated with requests to archive data. While "you would think that any [U.S.] snow and ice data would go into the [NSIDC], [4]." NSIDC is not funded for this. Instead funding is tied to specific programs that are driven by sponsor requirements. As a result, NSIDC as a whole receives many more requests for archival than can be accommodated. Moreover, the effort required to ingest a data set can vary dramatically both from data set to data set as well as with the level of service provided a given data set. For example, it takes much more effort to ingest an ongoing large satellite data stream, then it does to ingest a single, albeit large spreadsheet containing the measurements acquired during a field campaign. It also requires more effort to provide a variety of subsetting and reprojection services on a data set, than it does to simply make the data available from an anonymous ftp site. Consequently, processes that allow an archive to determine which data sets to acquire and what levels of service to provide the data sets accepted are becoming of ever increasing importance.

Weaver, et. al. [5] describe the considerations that NSIDC's Distributed Active Archive Center (DAAC) has been using when making data set prioritization decisions. These include considerations of data usage; stakeholder interest, in this case how relevant these data are to NASA and the NASA science community; product maturity, given by combining estimates of the data set's maturity as a science product, how preservable it is (i.e., its preservation maturity), and its likely societal impact; and the levels of service to be provided.

As the number of requests for archiving specifically aimed at the NSIDC DAAC increased, it became necessary for the DAAC to both standardize and codify the processes for assessing an archiving request in a manner that was simple to implement and which results in quantitative data set to data set comparisons. Section 3 describes the data acceptance process that resulted from these efforts, including methods for estimating likely data set activity levels, determining appropriate levels of service, and assessing product maturity. Section 4 summarizes the experiences and lessons learned over the first year of operations with these new procedures. A summary of areas for future work is given in Section 5.

REFERENCES

[1] CCSDS. 2002. "Reference Model for a Open Archival Information System (OAIS)." CCSDS 650.0-B-1. Blue Book. Issue 1. January 2002. [Equivalent to ISO 14721:2002]

[2] AGU. May 2009. “AGU Position Statement: The Importance of Long-Term Preservation and Accessibility of Geophysical Data.” http://www.agu.org/sci_pol/positions/geodata.shtml

[3] IPY Data Policy and Management Subcommittee. April 2008. “International Polar Year 2007-2008 Data Policy.” http://classic.ipy.org/Subcommittees/final_ipy_data_policy.pdf.

[4] B. Nelson. 9 September 2009. “Data sharing: Empty Archives.” Nature. 461, 218-223. doi:10.1038/461160a.

[5] Weaver, R. L. S., Meier, W. M., and R. M. Duerr. 2008. Maintaining data records: Practical decisions required for data set prioritization, preservation, and access. *Proceedings of the 2008 IEEE International Geoscience and Remote Sensing Symposium.* Volume 3, 617-619. doi: 10.1109/IGARSS.2008.4779423.