Designating User Communities for Scientific Data: Challenges and Solutions

Mark A. Parsons and Ruth Duerr
National Snow and Ice Data Center/World Data Center for Glaciology
Boulder, Colorado, USA
parsonsm@nsidc.org

Abstract

Explicitly defining a "designated user community" for a given data collection is essential to good scientific data stewardship. It enables data managers to determine what information needs to be developed and maintained to ensure the usability of the data now and into the future. A thorough understanding of the users helps managers define how to present and enable access to the data and may determine the actual format of the data. These considerations in turn have a direct impact on the long-term preservation of the data. Designating a specific user community is so important that it is explicitly called out as a mandatory responsibility of a an Open Archival Information System by the Consultative Committee for Space Data Systems in their ISO standard reference model. However, while defining a community may be essential, It is also extremely difficult, and it is impossible to predict how the use of a data collection may change over time. Indeed with today's rapidly changing information technology and scientific understanding, new data applications are likely to be discovered more frequently than ever. This creates a series of data management problems for data stewards be they traditional data archives or smaller nodes in a distributed or virtual data management system.

As data managers at the World Data Center for Glaciology, Boulder and the National Snow and Ice Data Center (NSIDC), we routinely confront and try to solve many of these problems. We certainly do not have all the answers, and it is likely that we will struggle with these problems forever. Yet we have begun to develop a set of best practices that mitigate these problems, and we believe these practices can be applied in a larger scientific context. We will use data collections held at NSIDC as case studies to illustrate this. We will also discuss how reexamining designated user communities might expand the use of science data.

1. Introduction

Scientific data preservation is pointless unless the data are used now and in the future. To ensure data usability, data managers need to understand *who* needs to use the data. The Consultative Committee for Space Data Systems (CCSDS) in its ISO standard Open Archival Information System (OAIS) reference model defines an archive as an "organization that intends to preserve information for access and use by a Designated Community" (CCSDS, 2002, p. 1-8). In principle, the data need to be "independently understandable" by the designated community. This principle can guide scientific data

mangers through decisions on data documentation, formats, and presentation, *if* they have a solid understanding of the designated community.

Scientists and archivists have long targeted their efforts toward specific communities. Journals are written for primary audiences. Libraries and archives target their content to specific user communities. Yet scientific data stewardship is a relatively new discipline that faces unique challenges in defining and meeting the needs of its communities. It can be difficult to define and understand a designated community for a scientific data collection because rapidly evolving technology and scientific understanding often lead to unanticipated users and applications for the data (Hunolt, 1999).

The National Snow and Ice Data Center (NSIDC) has been archiving and managing scientific data for more than 25 years. Although we have generally targeted our data to a relatively narrow cryospheric science community, we have seen a significant increase in the use of our data by an ever-broadening user base. In this paper, we use our experience to describe a set of principles and practices to help scientific data managers better define their user communities and improve data sets and related information to ensure long-term usability by both defined and potentially unanticipated users.

2. Defining a Designated Community

Often the initial user community for a data collection is self-evident. It is the participants in a particular experiment, the members of the science team for a remote sensing instrument, or a specific research group designated by a funding agency or scientific organization. Indeed, it is common for data access to be restricted to these narrow communities for a time, often to allow for instrument calibration and algorithm validation.

Once data are made more broadly available, it may seem reasonable to assume that the user community will be similar to the initial users. They will be educated in a closely related scientific discipline; they will be investigating similar phenomena; they will have a similar knowledge base. In short, they will understand. But, of course, this is a dangerous assumption. At a simple level, the new broader community may not understand the jargon particular to a certain experiment or mission. They will not be familiar with the assumptions, imperatives, and compromises that led to a particular data collection approach. They may not be as similar as originally thought. They will need more information, a greater context, to use the data effectively and accurately.

For example, the primary focus of the 2002-2003 Cold Land Processes Experiment (CLPX)[1] was to develop adequate understanding of snow pack characteristics over space and time to develop a satellite remote sensing method that could accurately measure the water content of snow (Cline, et al. 2003). As such, one might assume that users of these data would be reasonably versed in the engineering and physics of microwave remote

[1] See http://www.nohrsc.nws.gov/~cline/clp.html for an explanation of CLPX and http://nsidc.org/data/clpx for details of the data collected.

sensing. A review of our user registration logs suggests otherwise. Stated uses for the data included a variety of land surface modeling needs, atmospheric modeling, analysis of snow physics, permafrost studies, and, of course, remote sensing of snow. Moreover, most of the data has only been publicly available for a short time, so the user base is likely to continue to expand to include regional hydrologists, land and watershed managers, and others (a recent registrant plans to use the data to improve highway snow removal). So even though this started as a specialized and targeted data collection, making it publicly available made it necessary to recognize and support a broadly designated community comprised of many different user types.

In a more far-reaching example, the Special Sensor Microwave/Imager (SSM/I) has flown on U.S. Department of Defense polar orbiting satellites since 1987. The original purpose of the sensor was to support operational weather forecasting for the U.S. Air Force and Navy, but the data are used today for a broad array of global land, ocean, and atmospheric monitoring applications. Analysis of data from SSM/I and its predecessor the Scanning Multi-channel Microwave Radiometer has led to the creation of one of the longest satellite remote sensing time series of sea ice concentration and northern hemisphere snow cover, invaluable to global climate change studies. These applications are well beyond the scope envisioned by the original designers of the sensor. Other unanticipated uses continue today. For example, biologists have recently requested SSM/I derived snow water equivalent data to correlate with caribou calving dates.

The OAIS model emphasizes the need to consider a broad designated community early in the archiving process (CCSDS, 2002), but as the previous examples illustrate, it is unrealistic to think one can anticipate all uses of a scientific data collection. On the other hand, it is impractical to define an overly broad designated community such as the "general public," because it can be easily ill defined or even lead to inappropriate use of data (see Section 4). Nevertheless, data managers can prepare and maintain the data and documentation in a way that facilitates broad but appropriate use. (In this paper "documentation" includes all of what is commonly called metadata. The complete set of data and documentation is the Archive Information Package in OAIS vernacular)

3. Understanding Knowledge Bases —Conceptual Metaphors and Accepted Opinion

Philosophers have long grappled with the most effective way to convey information, but there is general agreement that you cannot even begin without an understanding of the underlying assumptions of both the information provider and recipient. Lakoff and Johnson (1980) argue that people need a conceptual basis to understand something and that scientists invoke key metaphorical concepts to work observations into a coherent, consistent structure. These metaphors may vary from discipline to discipline and are bound to change over time, even within a given discipline. Therefore, data managers need to be aware of the conceptual basis that led to the collection of a particular data set to effectively document its usability for a future community. Yet they need to be careful not to provide unnecessary detail. This is a central principle of classic rhetoric. Aristotle (1954, 1396a), wrote:

> … we must not carry [our] reasoning too far back, or the length of our argument will cause obscurity: nor must we put in all the steps that lead to our conclusion, or we shall waste words in saying what is manifest. It is this simplicity that makes the uneducated more effective than the educated when addressing popular audiences — makes them, as the poets tell us, "charm the crowd's ears more finely." Educated men lay down broad general principles; uneducated men argue from common knowledge and draw obvious conclusions. We must not, therefore, start from any and every accepted opinion, but only from those we have defined -- those accepted by our judges or by those whose authority they recognize.

The CCSDS (2002, p. 2-4) states "an OAIS must understand the Knowledge Base of its Designated Community to understand the minimum Representation Information that must be maintained." Although more narrowly defined by the CCSDS, this knowledge base should include Lakoff and Johnson's "conceptual metaphors" and Aristotle's "common knowledge" and "accepted opinion." Furthermore, it is equally important to understand the knowledge base of the data creator and how it coincides with or diverges from the knowledge base of the designated community. Finally, the data documentation must strike an appropriate balance between exposition and pith.

This is all rather abstract, but it is useful to have these principles in mind when developing data documentation. At NSIDC, the primary method we use to employ these principles is to have an educated, but non-expert, technical writer develop data set documentation in close consultation with scientists experienced with the type of data being described. We find it useful to collaborate closely with data providers when possible and we explicitly include staff scientists in our data management process. Others have emphasized the need for scientist involvement in data management and long-term archiving (Hunolt, 1999; Olsen et al., 1999), and we have found this involvement invaluable in developing documentation and supporting unanticipated users.

4. Characterizing Uncertainty

Writing data documentation for an unanticipated audience is a never-ending challenge, even when applying Aristotelian principles. Including certain elements in your document can help address that challenge. Chief among these required elements for scientific data is a frank and detailed explanation of the uncertainties of the data described. Scientific data creates a unique problem for data archival, in that there is always some degree uncertainty about the accuracy of the data. The uncertainty may be small with a simple in-situ measurement or relatively large with a satellite remote sensing time series or even unquantifiable with certain historical global measurements, but it is always there, and it must be well described to avoid data misuse. NSIDC data on sea ice concentration derived from passive microwave remote sensing provide a good case study of how to document uncertainty and why it is necessary.

NSIDC archives 38 data sets on sea ice and 12 derived from passive microwave remote sensing. This is a confusing array even for a sea ice specialist. Yet the diversity is largely due to the complexity involved in understanding a geophysical parameter in a region

where no one lives, few visit, and it is dark half the year. Passive microwave remote sensing has distinct advantages in this region since it does not require sunlight, it “sees” through clouds, and it measures over a broad area without requiring human visitation. Unfortunately, it has coarse resolution, is susceptible to contamination by weather systems, and has had problems with geolocation. Furthermore, since 1973 instruments and satellites have evolved making it difficult to construct a consistent time series or climate data record. In other words, there is a large uncertainty associated with these data sets.

NSIDC has sought to address this uncertainty in a variety of ways. First, we try and provide an overall context for the data. We have created a web site comparing and contrasting all the sea ice products we distribute (http://nsidc.org/data/seaice/). We describe the strengths and weaknesses of each product and indicate appropriate applications. The intent is to guide scientific users to the product that most suits their application. Once a user decides on a product, the data documentation includes detail written by both data providers and NSIDC scientists on the known errors and uncertainties in the data. In particular, we describe the different attributes of the two major algorithms used to generate sea ice concentration from passive microwave and changes in sensors over time. Some of this detail has required sophisticated scientific analysis that goes beyond the scope of typical data documentation. So in addition to citing the scientific literature, we have also developed a series of special reports on data related topics. Three of these reports are devoted to deeper analysis of passive-microwave-derived sea ice products (Maslanik et al., 1998; Stroeve et al. 1997; Stroeve and Smith, 2001). We hope we have adequately described relevant data uncertainty for future users and have organized it well enough to avoid obscurity.

All told, we are providing hundreds of pages of data documentation. Even with the guidance provided by our sea ice web site, it is unlikely we have met the needs of many non-expert users. The necessary information is available, but it is daunting. It is, however, necessarily daunting. With these products it is actually inappropriate to try and accommodate all uses. For example, say a wildlife biologist wanted to compare sea ice concentration with polar bear migration patterns. While these data are excellent for showing trends in sea ice concentration over time and over large areas, they are very inaccurate at the detailed spatial (pixel level) and temporal (daily) scale necessary for this application (Meier et al. 2001). For these products it is not only difficult, but also scientifically flawed, to define the designated community too broadly.

Instead we have derived specific products geared to appropriate non-expert use of passive microwave remotes sensing. Chief among these is our sea ice index (http://nsidc.org/data/seaice_index/), which provides images of average monthly ice conditions and trends and anomalies that compare recent conditions with the long-term averages (Figure 1). We specifically note the uses and limitations of the images. The idea is to provide an appropriate, high-level view of sea ice conditions and trends for people who are not experts in remote sensing of sea ice, be they other scientists or (to a lesser degree) the general public. For the general public, we also provide more general pages describing the role of sea ice in the global climate system (http://nsidc.org/cryosphere).

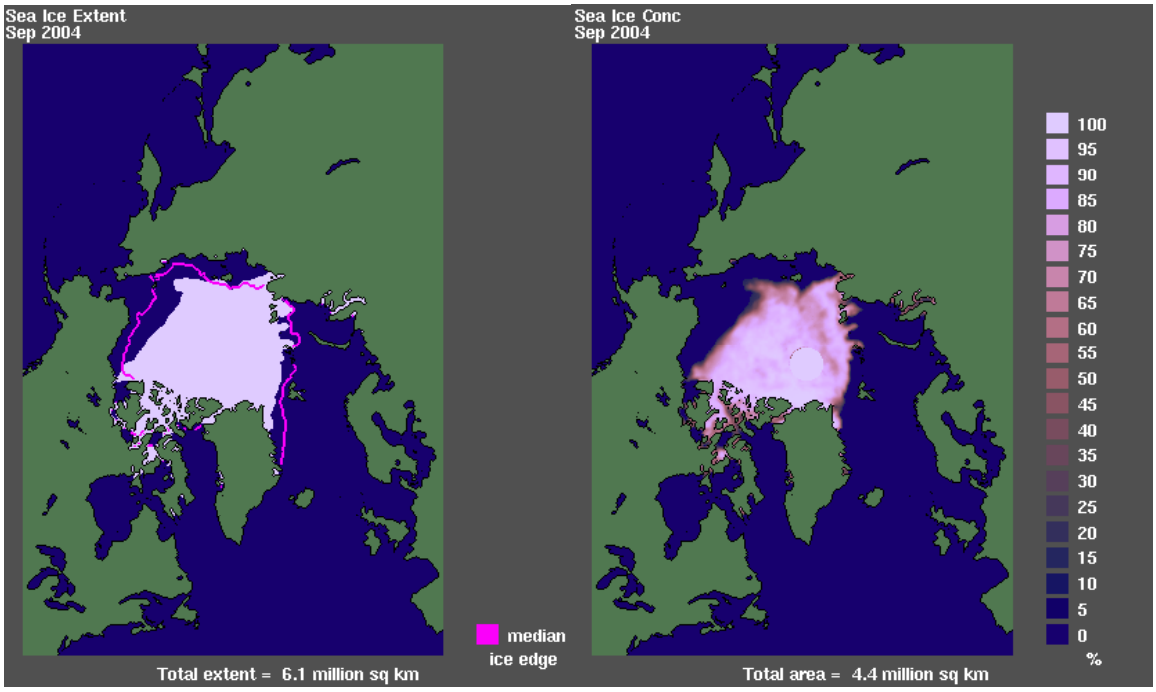

Figure 1. Examples of figures from NSIDC's Sea Ice Index showing Sept. 2004 ice extent compared to the median September extent for 1979-2000 (left) and September 2004 ice concentration (right) (Fetterer and Knowles, 2002).

Passive microwave remote sensing of sea ice is a complex business resulting in diverse and complex products. NSIDC's experience in presenting these products illustrates the need for scientific data managers to carefully document data uncertainty to facilitate broad and lasting data usability, but it also illustrates the importance of not defining too broad a designated community.

5. Determining Data Formats and Metaformats

So far, we have discussed documentation as a tool to enable broad and unanticipated use of scientific data. There are also attributes of the data itself that should be considered. Central among these is data format. Determining an appropriate data format for data storage and distribution is one of the more challenging problems of data management. Raymond (2004) argues that there are four important themes in designing file formats: interoperability, transparency, extensibility, and storage or transaction economy. He argues that the general file format that best addresses these themes is text, and that the only justification for a binary protocol is with very large data sets. Raymond is arguing from a computer programming perspective, but he takes a broad perspective and probably represents the generic, unanticipated user very well. His argument for text is compelling from a long-term archival perspective, especially in terms of its transparency, i.e. human readability. Nevertheless, some scientific data, notably high-volume remote sensing images, require a binary format for effective and efficient transfer and manipulation. When you do need to use a binary format, it is especially important to consider options to aid transparency such as example textual representations for some of the data.

The matter is further complicated when we begin to consider the specific implementation of a file format be it binary or textual. These specific implementations are what most generally consider the data format. Raymond (2004) calls them metaformats. Examples for text include delimiter-separated values and XML. Examples for binary include the hierarchical data format (HDF), GeoTIFF, and a variety of binary arrays. At a greater level of specificity is what the SEEDS Formulation Team (2003) calls the format profile. This is a specific implementation of a metaformat and would include machine-specific considerations such as byte order and 32 vs. 64-bit words. We will not discuss format profiles in great detail, but note there are few standards and they are subject to great variability and evolution in user requirements (SEEDS Formulation Team, 2003). If there is a broad user base, it would be ideal to distribute data in several formats and profiles generated on request and to preserve the data in yet another preservation-friendly format (Duerr et al., 2004).

The NASA Earth Observing System (EOS) experience illustrates the format issues. Very early in the program, NASA designated HDF-EOS as the standard format for all NASA EOS data. HDF-EOS was used both as the distribution and archive format. However, in response to the growth of the Geographic Information Systems (GIS), NASA plans to convert subsets of the data to GeoTIFF. In addition, NASA will provide the capability to reformat data on demand. So although these data are only a few years old, substantial effort has been invested to provide them in a format not originally anticipated as necessary.

Profiles can be even more complex. One profile worthy of special consideration in the Earth sciences is the map projection. Choosing an appropriate gridding and projection scheme is a perennial problem for data managers, much like choosing a data format. In contrast to data formats, though, it is often inappropriate to simply regrid data on demand, because the requested interpolation technique may not be appropriate for the data and spatial resolutions of source and target grids. There is no single “best” projection for all applications (Knowles, 1993). We believe that one solution is to clearly define the necessary grid or grids in the experiment design and sample accordingly, while at the same time provide the swath data for users who may require it. This approach can, however, increase storage requirements by orders of magnitude. More sophisticated solutions such as storing remote sensing data in a geodatabase using spherical (or geoidal) geometry should also be considered.

6. Conclusion

Good scientific data stewardship requires explicit recognition and understanding of data user communities. It is often necessary to define a broad user community and still recognize that there will be unanticipated uses of a given scientific data collection. At the same time, one should recognize the risk of scientifically inappropriate use of data and be sure to carefully document data uncertainty. It may also be necessary to devise creative methods of data presentation to ensure broad but appropriate use. This requires data stewards to challenge their basic assumptions about what is understood about their data both by data users and providers, to ensure comprehensive documentation, and to provide data in transparent, interoperable formats. Ultimately, these fundamental long-term

archive considerations may be summarized with the basic principle — “keep things simple but flexible”.

Industrial designers, software engineers, and even traditional archives and libraries rely heavily on studies of the use of their systems. The results of some of these studies may help guide data managers in the development of data access systems, but they provide little if any guidance on the necessary attributes of the actual data and supporting documentation. The value of scientific data increases with use. Therefore, data stewards need to better understand who will use their data and what they can do to facilitate that use. This is vital research area for the long-term preservation of scientific data.

7. References

Aristotle. 1954. *Rhetoric*. translated by W. Rhys Roberts. New York: Modern Library. 289 p.

CCSDS (Consultative Committee for Space Data Systems). 2002. *Reference Model for an Open Archival Information System (OAIS) CCDSD 650.0-B-1. Blue Book. Issue 1*. Washington, DC: CCSDS Secretariat. 148 p. [Equivalent to ISO 14721:2002].

Cline D., K. Elder., R. E. Davis, J. Hardy, G.E. Liston, D. Imel, S.H. Yueh, A.J. Gasiewski, G. Koh, R.L. Armstrong, and M.A. Parsons. 2003. “Overview of the NASA Cold Land Processes Field Experiment (CLPX-2002).” *Proc. SPIE Int. Soc. Opt. Eng*. 4894: 361-372.

Duerr R., M.A. Parsons, M. Marquis, R. Dichtl, and T. Mullins. 2004, “Challenges in long-term data stewardship”. *Proc. 21st IEEE Conference on Mass Storage Systems and Technologies.* NASA/CP-2004-212750: 47-67.

Fetterer, F. and K. Knowles. 2002, updated 2004. *Sea Ice Index*. Boulder, CO: National Snow and Ice Data Center. Digital media.

Hunolt, G., compiler. 1999. *Global Change Science Requirements for Long-Term Archiving. Report of the Workshop, Oct 28-30, 1998*. Washington, DC: USGCRP Program Office. 56 p.

Knowles K.W. 1993. “Points, pixels, grids, and cells – a mapping and gridding primer.” http://cires.colorado.edu/~knowlesk/ppgc.html. Accessed 2 September 2004.

Lakoff, G. and M. Johnson. 1980. *Metaphors We Live By*. University of Chicago Press 242 p.

Maslanik, J., T. Agnew, M. Drinkwater, W. Emery, C. Fowler, R. Kwok, and A. Liu. 1998. *Summary of Ice-Motion Mapping Using Passive Microwave Data. NSIDC.* Special Report 8. Boulder, CO: National Snow and Ice Data Center. http://nsidc.org/pubs/special/nsidc_special_report_8.pdf. Accessed 22 October 2004.

Meier, W.N., M.L.V. Woert, and C. Bertoia. 2001. Evaluation of operational SSM/I algorithms. *Annals of Glaciology* 33: 102-108.

Raymond, E. S. 2004 *The Art of UNIX Programming.* Boston, MA: Addison-Wesley. 525 p.

SEEDS Formulation Team. 2003. Strategic Evolution of Earth Science Enterprise Data Systems (SEEDS) Formulation Team final recommendations report. http://lennier.gsfc.nasa.gov/seeds/FinRec.htm. Accessed January 2004.

Stroeve, J, X. Li, and J. Maslanik. 1997. *An Intercomparison of DMSP F11- and F13-derived Sea Ice Products.* NSIDC Special Report 5. Boulder, CO: National Snow and Ice Data Center. http://nsidc.org/pubs/special/5/index.html. Accessed 22 October 2004.

Stroeve J. and J. Smith. 2001. *Comparison of Near Real-Time DMSP SSM/I Daily Polar Gridded Products and SSM/I Polar Gridded Products.* NSIDC Special Report 10. Boulder, CO: National Snow and Ice Data Center. http://nsidc.org/pubs/special/nsidc_special_report_10.pdf. Accessed 22 October 2004.

Olson R.J., J.M. Briggs, J.H. Porter, G.R. Mah, and S.G. Stafford. 1999. "Managing data from multiple disciplines, scales, and sites to support synthesis and modeling." *Remote Sensing of Environment.* 70**:** 99-107. DOI: 10.1016/S0034-4257(99)00060-7