

Mark A. Parsons



National Snow and Ice Data Center/World Data Center for Glaciology, Boulder

parsonsm@nsidc.org

Mark Parsons is a program manager at the National Snow and Ice Data Center (NSIDC) and World Data Center for Glaciology, Boulder. He has been involved in data management for nearly 20 years, during which he defined and implemented the overall data management process for NSIDC. He is active in multiple international informatics efforts and co-chairs the International Polar Year (IPY) Data Management and Policy Subcommittee. He is also leading an effort with many international partners to host the IPY Data and Information Service, endorsed by the ICSU/WMO Joint Committee for IPY.

GRL2020 Position Paper

I write this position paper from three related perspectives. First, I write from the perspective of a long-term manager at the World Data Center for Glaciology, Boulder, which began in 1957 as a library of documents and photographs from the International Geophysical Year. The data center has grown to be a national data center archiving large remote-sensing and other digital collections, yet we still maintain an active and growing library with the original documents and photos. Second, I write from the perspective of what is needed for the current International Polar Year 2007-2009 (IPY). IPY is a large, international, and interdisciplinary science program that has been running for two years—a \$430 million addition to existing polar research involving 63 nations addressing all physical, life, and social science disciplines related to Earth science in one region. As such, IPY provides an excellent test case for data management issues for all of Earth science. Finally, I write from the perspective IPY Data and Information Service (IPYDIS). The IPYDIS is a global partnership of data centers, libraries, archives, and networks working to ensure proper stewardship of IPY and related data, but I will focus on the role of libraries.

The central requirement of the IPYDIS is to provide data to scientists to enable data based research. It seeks to 1) identify what data are collected as part of IPY; 2) serve the data through interoperable protocols; and 3) preserve the data in accordance with the Open Archival Information System (OAIS) Reference Model (CCSDS 2002) at libraries and data centers around the world.

When identifying the data, we should consider the three basic categories of digital data defined by the National Science Board—reference data, resource or community data, and research data (NSB 2005)—and how these different categories of data create different issues and policy implications. Reference data are usually well managed and readily available. Community data are highly variable depending on the community, but a primary issue is making those data available outside the original community. Research data are the most diverse and are often the hardest to locate, especially in the very interdisciplinary, grass roots effort of IPY. Often the only way to identify a data collection, let alone ensure its access and preservation, is by direct contact with individual investigators. This is an area libraries could provide excellent service. Through existing close connections to their institutional communities, libraries could help investigators describe and

deposit their data into open repositories in a timely manner¹. More importantly, libraries are often well situated to educate the current and next generation of scientists on effective data management practices.

Once data are identified, they must be served to the broader community. Libraries, community-based archives, and national and international data centers need to partner to readily share data and metadata in a distributed environment. To facilitate metadata sharing, IPY has developed a basic metadata profile that also provides a crosswalk across several standards. Based on this profile, we are beginning to develop a union catalog of shared metadata using OAI-PMH and other XML-based protocols. Extending this to the library community is an obvious next step. We need to address technical challenges such as de-conflicting records, but a more significant challenge is coordination across diverse disciplines that use different standards, processes, vocabularies, metaphors, and assumptions. Further, interdisciplinary data access systems are addressing new user communities. We need to carefully consider what the OAIS Reference Model calls the “designated community” for a given collection, because this, in turn, defines many data archival and access requirements. This is especially important when we consider IPY’s explicit involvement of Arctic residents as both a designated user community and as a data providers through community-based monitoring and their own local and traditional knowledge (Allison et al. 2007, ICSU 2004a). Further, we must recognize that user communities can change over time, often in unanticipated ways (Parsons and Duerr 2005). Some efforts have been made to develop more sophisticated semantics and ontologies, but it is a young field, ripe for more library involvement.

We should also consider how these user communities think. For example, David Fulkner, in a keynote presentation² at an Arctic research conference, showed how scientists have two worldviews. One view sees the world as a collection of features arranged in space (e.g. GIS users), while the other view sees the world as a set of parameters that vary over time (e.g., climate modelers). While Fulkner emphasizes that this is an over-simplified dichotomy, it illustrates how the two basic approaches to data integration (i.e., integration through time or space) may be relevant in different situations. The primary data sharing services (outside of basic FTP and HTTP) emerging as community standards in IPY (THREDDS and OGC³) also reflect this dichotomy (although there is increasing convergence). Again, libraries and data centers need to collaborate in understanding and capturing the experience of diverse, evolving communities.

Finally, the most important issue is to preserve the data for the long-term. Preservation may be the best understood problem but also the most intractable. We know *what* to do, but we struggle with *how*. The OAIS model provides a good baseline of what is required, but detailed implementation is only beginning to be defined for the Earth sciences and it is clearly complex and costly (cf. Duerr, et al. 2007). This may be the most critical and fruitful area for library involvement. Not only do libraries have centuries of experience in data preservation and *curation*, but they also have the necessary respect and support of institutions and society to provide long-term sustainability. A core challenge is to develop a sustainable business model whereby the entire scientific community and society at large contribute to the sustained preservation of unique and critical data in an era of rapid technological, social, and environmental change.

Ultimately, my vision is of a data preservation and access “utility”— a core infrastructure of science that is simple, predictable, reliable, extensible, accessible, and durable. But just like with existing utilities, such as water, electricity, and communications, the basic simplicity on the surface belies deep complexity, structure, planning, and professionalism. Creating that level of infrastructure requires great collaboration around standards, maintenance, and professional development and certification. We must bridge cultural barriers between scientific disciplines, between data managers and researchers, between libraries and data centers.

The time is ripe for greater interdisciplinary collaboration and coordinated Earth system informatics, as is evident in initiatives such as IPY, the Electronic Geophysical Year (eGY), the Global Earth

¹ IPY Data Policy states that data should be available “freely, openly, and on the shortest feasible timescale”. The requirement of timely data release may be one of the most controversial. See http://classic.ipy.org/Subcommittees/final_ipy_data_policy.pdf

² <http://www.eol.ucar.edu/projects/aon-cadis/meetings/200703/misc/Fulker/>

³ Open Geospatial Consortium: <http://www.opengeospatial.org/standards>
Thematic Realtime Environmental Distributed Data Services: <http://www.unidata.ucar.edu/projects/THREDDS/>

Observing System of Systems (GEOSS), the International Council of Science's goal to take a leadership role in data management for the sciences (ICSU 2004), a proposed Union Commission for Data and Information in the geosciences, and this Global Research Libraries 2020 conference. In particular, GRL2020 could work to build greater collaboration and information sharing, even convergence, between libraries and data centers. Under the auspices of eGY, Rajendra Bose, I, and others will propose a session at the American Geophysical Union fall meeting tentatively titled the "Evolution of Research Libraries and Data Centers in Earth and Space Sciences" to begin a formal discussion on library-data center collaboration, and potentially leading to refereed publications in the growing body of informatics literature. Other disciplines may want to consider similar initiatives. Other opportunities for collaboration include increased metadata sharing between libraries and data centers through OAI-PMH and related protocols in organized consortia, and the development of formal data management curricula and education programs.

References Cited

- Allison, I, M Belánd, K Alverson, R Bell, D Carlson, K Danell, C Ellis-Evans, *et al.*. 2007. *The Scope of Science for the International Polar Year 2007–2008, WMO/TD–No. 1364*. Geneva: World Meteorological Organization.
- CCSDS (Consultative Committee for Space Data Systems). 2002. *Reference Model for an Open Archival Information System (OAIS) CCSDS 650.0-B-1 Issue 1*. Washington, DC: CCSDS Secretariat.
- Duerr, R, R Weaver, and M Parsons. 2006. A New approach to preservation metadata for scientific data: A real world example. *Geoscience and Remote Sensing Symposium, 2006 IGARSS 2006 IEEE International Conference on*. pp. 305-308.
- ICSU (International Council for Science). 2004. A Framework for the International Polar Year 2007-2008.
- ICSU (International Council for Science). 2004. ICSU Report of the CSPR Assessment Panel on Scientific Data and Information
- NSB (National Science Board). 2005. Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century National Science Foundation. 87 pp.
- Parsons, MA, and R Duerr. 2005. Designating user communities for scientific data: challenges and solutions. *Data Science Journal*. 4:31-38.

