Interdisciplinary data discovery and use—The Arctic experience

The Arctic is in the midst of rapid change in its land, ocean, atmospheric, biologic, and social systems. Recent decades have seen pronounced rises in surface air temperature, attended by warming and thawing of permafrost, transitions from tundra to shrub vegetation, and strong downwards trends in sea ice volume and summer sea ice extent. There is evidence of shifts in atmospheric and oceanic circulations. Events unfolding in the Arctic are having significant social impact regionally and may have hemispheric and even global-scale ramifications (ACIA, 2005; Serreze et al., 2000; Serreze et al., 2007; Sommerkorn, 2009). In recent years, major initiatives such as the US Arctic Observing Network (AON) and the International Polar Year (IPY) have sought to understand the Arctic holistically as a system. Correspondingly, both initiatives emphasized the need for interdisciplinary and integrative data management much in line with the EarthCube vision (ICSU, 2004; NRC, 2006). AON even laid out an ambitious and visionary use case (see box).

Imaginary Journey Through the Arctic Observing Network Portal

Suppose one day you read an interesting article speculating on the contribution of processes in submarine canyons to the global carbon cycle and decide to explore arctic datasets. Entering the AON data portal, you first encounter icons for terrestrial, atmospheric, oceanic, and human dimensions that contain a summary of data holdings under each discipline. You then have the option to browse datasets by discipline or by theme. Using data exploration tools, you search for canyon processes and determine what relevant meteorological, geophysical, and oceanic datasets are archived, and their availability in space and time.

Although you do not realize it, the information accessed comes from four different data centers in two different countries. For observations that are interesting but unfamiliar, you find links to descriptions of the instrumentation, the methods, and the data processing steps. You also find links to browse images of the datasets and, after inspection of these, you decide flow levels of the X River bear closer investigation, as the X River appears to be associated with the Y Canyon, and both the oceanic and terrestrial environments are well instrumented.

Plotting the time series using the online data display tools, you observe that three years ago, in June, the gauges reported an abrupt drop in water level after a gradual rise through late spring. The screen also shows an icon that looks like the silhouette of a parka. Curious, you click on the icon, and a text box pops up describing a large ice dam that gave way about the time of the abrupt water level drop, with a notation from the Inuit hunter who reported the event. Now you open the relational database interface in the AON portal and frame a query requesting turbidity measurements within 100 km of the mouth of the X River during the timeframe of the ice dam collapse. Within seconds, you have links to data streams and generate another series of plots. These show an increase in turbidity within the Y Canyon two days after the ice dam collapsed. You suspect that you have identified a flow event carrying sediment into the deep Arctic Ocean. Wondering how general these events are, you search for abrupt drops in tide gauge measurements coupled to local increases turbidity measurements for other arctic river systems and find three more candidate events.

It is almost the end of the day and you download your time-series plots and email them to your colleagues twelve time zones away for their review tomorrow. You save your AON session using the password protection you have installed so that you can access the data again tomorrow without having to redo the data searches. Before wrapping up, you post a request to the event detection service, providing the combined tide gauge turbidity criteria as the trigger. Finally, you post a request to the observation scheduling list, starting the process to request time on the docked autonomous underwater vehicle near the mouth of the X River to be triggered on detection of an event. It has been a productive day.

— *Toward an Integrated Arctic Observing Network.* (NRC, 2006 p. 41)

Several years ago, NSF supported the Cooperative Arctic Data and Information System (CADIS), a collaboration between labs at UCAR, Unidata, and the National Snow and Ice Data Center (NSIDC) at the University of Colorado, to help achieve this vision of interdisciplinary data discovery, access, and use. NSF also supported the Exchange for Local Observations and Knowledge in the Arctic (ELOKA) to facilitate the collection, preservation, sharing, and use of local and traditional knowledge. Both of these efforts were major contributors to data coordination and sharing efforts under the IPY.

Not surprisingly, we have found interdisciplinary data discovery, let alone full integration, to be extremely challenging. We have found that much of the emphasis of data-intensive science or e-science has been on dealing with overwhelming data volumes (e.g., Hey et al., 2009), and has underplayed another equally daunting challenge—the *diversity* of interdisciplinary data and the need to interrelate these data to understand complex systemic problems such as environmental change and its impact.

A major challenge of interdisciplinary, systemic research is in the use of so-called research collections. The National Science Board (NSB) defines three broad categories of data collections: research, resource, and reference collections. Research collections, which are data collected by individual investigators and small research groups, stand out as a critically absent from current data systems. These unique observations of the Earth system are unrepeatable and increase in value over time, but they are underutilized, vanishing, and often forgotten (NSB, 2005). Heidorn (2008) describes these data as the long tail of science—the small, but very diverse data sets collected by the majority of scientists, and the heart of NSF's data portfolio. It is this heterogeneity of data and the cultural diversity of their collectors, not their volume, that present the greatest challenges. Furthermore, these research data need to be integrated with reference and community data, and that requires adaptation by both data managers and collectors. As data managers for IPY, we found that while technology is a critical factor to addressing the challenges of data diversity, the technologies developing for exa-scale data volumes are not the same as what is needed for extremely distributed and heterogeneous data. Furthermore, as with any sociotechnical change, the greater challenges are more socio-cultural than technical.

While grand challenges remain, we believe several lessons from our recent hands-on experience managing Arctic data for synthesis could inform the EarthCube effort both theoretically and practically:

- IPY and the NSF AON program both had forward-looking, *open but ethical*, data sharing policies. This provided a solid foundation for collaboration around data. Furthermore, both initiatives had lightweight, grass-roots style governance mechanisms. We found this essential for very diverse communities to collaborate. It is in keeping with the governance concepts laid out in the ESIP and Unidata governance whitepapers.
- Different disciplinary communities have different expectations, tools, analytical needs, data sharing cultures, conceptual metaphors, and knowledge bases (Key Perspectives Ltd, 2010; Parsons and Duerr, 2005; Parsons et al., 2011a). One size does not fit all, and successful projects both lead and follow. They are responsive to community needs while also driving new ideas. They also understand user workflows and try to accommodate to those workflows.
- We need to recognize that 1) data will be highly distributed and housed at many different types of institutions; 2) the use and users of the data will be very diverse and even unpredictable; and 3) the types, formats, units, contexts and vocabularies of the data will continue to be very complex if not chaotic. In this context, the "one-stop shop" is a poor data discovery metaphor. It is better to think of a marketplace or bazaar—a virtual space where all data can be found, but specialist portals provide the expertise, information, and referrals necessary to identify and understand data within a specific disciplinary context. It is necessary for technologies to be lightweight, flexible, and service based. We should be building an Earth science development

platform that allows different communities to enhance or adapt the system to meet their needs[1]. This requires the sort of service-oriented approaches described in the ESIP and GI-Cat whitepapers. Centralized registries of data and services will not fully scale to handle the diversity. Data managers need to employ more open, cloud-based approaches of data broadcasting and customized aggregation, and they need to make their data and metadata available through a variety of protocols using multiple formats to serve variable communities. Systems need to start simple and iterate to expand their interconnection with other systems and user communities over time.

- Several projects in IPY, especially CADIS, found in working with the data providers that it is a worthy and perhaps necessary long-term goal to ask all investigators to submit their data using certain standards and conventions, but this is not always feasible. The necessary education and transformation will take considerable time and collaborative effort. These and other collaborative and consensus-based decision making efforts require time, commitment and sustained, well-sponsored effort.
- We have found that a pragmatic, flexible, and especially a *collaborative* approach to system development can encourage the behavioral adaptation in science that enhances data flow and usefulness. Collaboration between data managers and data creators can have broad, positive repercussions. At one level, congenial collaboration and the creation of simple data sharing tools and cookbooks can increase the use of standards and improve data and metadata completeness. At a deeper level, Saracevic suggested that "the *subject knowledge view of relevance is fundamental to all other views of relevance,* because subject knowledge is fundamental to communication of knowledge." (1975 p. 333 his emphasis). In other words, we cannot communicate the value, uncertainty, or usability of data effectively without capturing expert knowledge of the data creator. This suggests that while it is important to consider user needs and perspectives, it is equally important to consider data creator perspectives in order to effectively capture contextual knowledge. Data can serve as the sort of "boundary objects" central for translating between viewpoints in the interdisciplinary discussion of science, but the objects must remain robust enough to maintain identity across disciplines (Star and Griesemer, 1989). Metadata is best generated through collaboration between the subject- knowledge holder and information specialists. This makes the data more useful within and without specialist communities.

Finally, we must consider the data "ecosystem" and recognize that informatics solutions are rapidly evolving and selection processes are going on all the time. These solutions are not usually the result of any central plan, and they require continued organic adaptation by technology, people, organizations, and society. This adaptability must be a central tenant of EarthCube and the Arctic data and research community within the US and around the globe is eager to participate.

> *— Mark A. Parsons on behalf of NSIDC, the CADIS and ELOKA teams, and the general Arctic data community. Much of the document was cribbed from Parsons et al. (2011b).*

[1] This concept of building a "platform" for science rather than tools or products was partially inspired by this insightful post from Google developer, Steve Yegge:
https://plus.google.com/112678702228711889851/posts/eVeouesvaVX

References

ACIA. 2005. *Impacts of a Warming Arctic, Arctic Climate Impacts Assessment (ACIA).* Cambridge University Press.

Heidorn, PB. 2008. Shedding light on the dark data in the long tail of science. *Library Trends. 57*(2):280-299. http://dx.doi.org/10.1353/lib.0.0036

Hey T, S Tansley, and K Tolle (ed.).2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* USA: Microsoft Research.

ICSU. 2004. *A Framework for the International Polar Year 2007-2008.* Paris: International Council for Science.

Key Perspectives Ltd. 2010. *Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long Term Viability.* Edinburgh: Digital Curation Center. http://www.dcc.ac.uk/sites/default/files/SCARP%20SYNTHESIS_FINAL.pdf. Accessed 5 Feb. 2011.

National Research Council. 2006. *Toward an Integrated Arctic Observing Network.* Washington, DC: National Academies Press. http://www.nap.edu/catalog/11607.html. Accessed 14 Oct. 2011.

NSB (National Science Board). 2005. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century.* Washington, DC: 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.

Parsons, MA, T de Bruin, S Tomlinson, H Campbell, Ø Godøy, J LeClert, and IPY Data Policy and Management SubCommittee. 2011a. The state of polar data—the IPY experience. In I Krupnik, I Allison, R Bell, P Cutler, D Hik, J López-Martínez, V Rachold, E Sarukhanian, and Summerhayes (ed.). *Understanding Earth's Polar Challenges: International Polar Year 2007-2008* Edmonton, Canada: CCI Press. pp. 457-476.

Parsons, MA, Ø Godøy, E LeDrew, TF de Bruin, B Danis, S Tomlinson, and D Carlson. 2011b (in press). A conceptual framework for managing very diverse data for complex interdisciplinary science. *Journal of Information Science. http://dl.dropbox.com/u/546900/JIS-1391-v6.pdf*.

Saracevic, T. 1975. RELEVANCE: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science. 26*(6):321-343. http://dx.doi.org/10.1002/asi.4630260604

Serreze, MC, JE Walsh, FS Chapin, T Osterkamp, M Dyurgerov, V Romanovsky, WC Oechel, J Morison, T Zhang, and RG Barry. 2000. Observational evidence of recent change in the northern high-latitude environment. *Climatic Change. 46*:159-207.

Serreze, MC, MM Holland, and J Stroeve. 2007. Perspectives on the Arctic's shrinking sea-ice cover. *Science. 315*(5818):1533-36. http://dx.doi.org/10.1126/science.1139426

Sommerkorn M, and SJ Hassol (ed.).2009. *Arctic Climate Feedbacks: Global Implications.* WWF Internationnal Arctic Programme.

http://www.worldwildlife.org/what/wherewework/arctic/WWFBinaryitem13543.pdf. Accessed 14 Oct. 2011.

Star, SL, and JR Griesemer. 1989. Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science. 19*(3):387-420. http://www.jstor.org/stable/285080