LEARNING FROM THE INTERNATIONAL POLAR YEAR TO BUILD THE FUTURE OF POLAR DATA MANAGEMENT

M Mokrane[1]*, M A Parsons[2]

[1]ICSU World Data System IPO, c/o NICT, 4-2-1 Nukui-kitamachi, Koganei, 184-8795 Tokyo, Japan
**Email:* mustapha.mokrane@icsu-wds.org
[2]Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY, 12180, USA
Email: parsom3@rpi.edu

ABSTRACT

The research data landscape of the last International Polar Year was dramatically different from its predecessors. Data scientists documented lessons learned about management of large, diverse, and interdisciplinary datasets to inform future development and practices. Improved, iterative, and adaptive data curation and system development methods to address these challenges will be facilitated by building collaborations locally and globally across the 'data ecosystem', thus, shaping and sustaining an international data infrastructure to fulfil modern scientific needs and societal expectations. International coordination is necessary to achieve convergence between domain-specific data systems and hence enable multidisciplinary approaches needed to solve the Global Challenges.

Keywords: International Polar Year, Data management, Data curation and stewardship, Long-term preservation, Open-access, Data infrastructure, Data ecosystem

1 INTRODUCTION

Coordination of international polar data management activities benefitted greatly from the burst of research activities generated by the International Polar Year 2007–2008 (IPY). Since the end of IPY and its dedicated data management activities, however, polar data management has improved at a relatively slow pace right when data sharing, reuse, and interoperability, and the sustainability of eInfrastructures are increasingly recognized as important by senior science funders and policymakers (G8+O5 Global Research Infrastructure Sub Group on Data, 2011).

IPY was built on the successful past models of the International Geophysical Year (IGY) in 1957–1958 and even the original IPY in 1882–1883 (Rapley, Bell, Allison, Bindschadler, Casassa, Chown, et al., 2004). IGY was a good example of a successful data-centric and internationally coordinated research programme. One of its lasting successes, and possibly only institutional legacy, was the World Data Centres (WDCs) established by the International Council for Science (ICSU), the leading nongovernmental global scientific organization, as the first internationally coordinated effort to preserve and make scientific data openly and freely accessible. By fulfilling their mandate for over half a century, WDCs effectively set the standard for Open Access to scientific research data and influenced the global data management landscape (Aronova, Baker, & Oreskes, 2010).

When seen from a purely scientific perspective, IPY, and its predecessor models, were undoubtedly successful, multidisciplinary research endeavours. At the same time, they revealed challenges facing the international scientific community to coordinate management, preservation, and dissemination of scientific data (Carlson, 2011), particularly of the diverse research collections from the so-called 'long-tail of science'. IPY was a very large and complex project, with an estimated budget of 1.2 billion USD and approximately 50,000 participants from 63 nations (Carlson, 2010), that presented daunting data stewardship challenges to the polar research community. Soon after IPY, many data scientists attempted to document lessons learned about stewardship of complex, sometimes large, diverse, and interdisciplinary data. Several reports were produced by national agencies and international organizations. In particular, the IPY Data Committee and the IPY Data and Information Service (IPYDIS) conducted two major analyses of IPY data management (Parsons, Godøy, LeDrew, de Bruin, Danis, Tomlinson, et al., 2011; Parsons, de Bruin, Tomlinson, Campbell, Godøy, & LeClert, 2011). These reports all converged in their analysis and recommended direct involvement of data scientists at every level, from senior management to field and laboratory support, in the early planning and throughout the execution of research programmes, implying also that funding data activities must be an integral part of the scientific research effort.

Now, nearly five years after the end of IPY, we attempt to examine the lasting lessons of IPY and propose solutions to help enable a global framework for international polar data management.

2 IPY DATA INFRASTRUCTURE

Because data resulting from IPY were seen as potentially the most important single outcome of the programme, its planners laid out a noble and ambitious data management plan (ICSU, 2004). An IPY Data Committee developed a visionary data policy, and polar data scientists around the world rallied to form the distributed IPYDIS. Polar data management policy and practice advanced immensely, but few would say that IPY has met the vision and all of the objectives originally planned. As is often the case, a critical concern during the initial phase was the lack of adequate funding for data management and international coordination. This continues to be a concern for data management in polar research projects in general but was perhaps not the core issue. Instead, the way the community approached the challenges of truly interdisciplinary data sharing was somewhat naïve.

It was assumed that creating a data service from existing components and infrastructures was enough, following the system of systems model popularized by the Global Earth Observation System of Systems (GEOS, Battrick, 2005). Retrospectively, we now propose that IPY, and polar data management in general, needed a more diverse, dynamic, scale-free, and adaptive data system that was reliant on multiple social and technical components to form an entirely renewed data ecosystem.

The main lesson derived from the IPY data stewardship experience is that building appropriate data infrastructure to enable international sharing and reuse of multidisciplinary datasets is a complex and fraught sociotechnical exercise. It surely requires sustainable funding, but more importantly, it requires time, patience, and a highly adaptive and creative community effort. We describe four overarching themes that can inform the overall process and that guide specific data stewardship activities supporting the development of data infrastructure.

2.1 The challenge in diversity

In IPY, we found that the greatest data stewardship challenge lies in the diversity of all the data necessary to understand complex systems such as the polar regions. Furthermore, research collections are central to polar research, yet they can be highly disparate and challenging to manage.

Different disciplines have different data systems at various levels of maturity as well as different attitudes and norms of behaviour around data sharing, all of which affect how we build integrated systems. For example, centralized metadata registries become unwieldy and imprecise when describing heterogeneous objects to potentially diverse audiences. Instead, a federation of specialized data systems and portals using open web services is preferable, a data ‘bazaar’ rather than a ‘one-stop shop’.

2.2 Communities and collaboration

Interoperability, indeed infrastructure, is built through relationships. The tacit knowledge of specialization is revealed and shared through relationships, and these are the foundation upon which to build a collaborative community. It was found during IPY that relationships both between different data scientists and amongst data scientists, users, and providers improved data systems, documentation, and the data themselves. Great value was found in creating a global polar data community while also integrating data scientists into their local disciplinary communities. Data scientists are often ‘in between’ workers or intermediaries who can help build community. Improving data scientists’ training and career development, especially at early stages, is fundamental to nurturing and improving the global polar data community. For example, the Association of Polar Early Career Scientists, an IPY-offshoot organization that facilitates networking and promotes education and outreach for undergraduate and graduate students (Baeseman & Pope, 2011), plays a key role in building the polar data community and must be strengthened.

2.3 Methods and training

Part of data scientists’ training needs to include instruction on methods and improving relationships and collaboration. We learned that when developing data systems, the best method is to start simple, using proven approaches, and then take an incremental, iterative approach to expanding their interconnection. This means that system designers need to work closely with, and be responsive and adaptive to, both data providers and users. Furthermore, user expectations and needs change over time, and systems need to continuously evolve for

optimal capability. This requires more than use-case-driven, agile development; it also requires case studies and ethnographic and cognitive science approaches to understand how people conceive, produce, and use data.

2.4 Globalism and localism

Infrastructure works across all scales. It must function locally and reach globally. It is important to be constantly building relationships both globally and locally, to act 'glocaly'. For example, the real impact of the IPY data policy was felt when it was enforced by national governments, but the international recognition of the policy led national governments to act. Correspondingly, a union catalogue of IPY datasets could not begin to be built until local data centres were established and functional. In some cases, it took years of cultivating local partnerships before they could extend more broadly.

Regional success contributes to global success, which pushes local success. The polar community should continue to foster its own polar and disciplinary-orientated communities while participating in global initiatives such as the ICSU World Data System (ICSU-WDS), a network of multidisciplinary data centres and data services established by ICSU, and the Research Data Alliance (RDA), an international community effort to improve data sharing.

3 EVOLUTION OF GLOBAL DATA SERVICES

Important lessons derived from the IPY experience influenced the strategies of many international organizations, which have consequently started new, or adapted existing, initiatives to improve sharing and reuse of scientific research data.

For example, ICSU launched its World Data System in 2009 to reform and build upon the legacy of its former World Data Centres and Federation of Astronomical and Geophysical Data Analysis Services. These bodies were not able to respond in a coordinated way and fulfil the data needs of IPY. In particular, there were no mechanisms in place to cater for the diverse datasets of the 'long-tail of science' and thus meet the high expectations of the IPY designers. To address these deficiencies and to prepare an effective response to the coming challenges of other major programmes, such as the ICSU-sponsored Future Earth initiative (Future Earth Transition Team, 2013), the new organization is striving to build worldwide 'communities of excellence' for scientific data services (Harris, 2012). To achieve this goal, its Scientific Committee has identified at least three pillars to build upon. The first pillar is establishing the *trustworthiness* necessary to enable interoperability at the technical and social levels. It is achieved, at least partially, by certifying Member Organizations, holders and providers of data or data services, using internationally recognized standards, and ICSU-WDS is taking the lead in this area. The second pillar is *stewardship* to improve data discovery, data preservation, and reusability; ICSU-WDS is working with its Member Organizations and partners to realize searchable, interoperable, and distributed common infrastructure. The third pillar is *inclusiveness*, both in geographical and disciplinary coverage. Active recruitment of Member organizations in the Social Sciences and Humanities has led to a visible expansion of ICSU-WDS in these domains. WDS geographical coverage has also noticeably improved compared with its predecessor bodies, including through committed nurturing of initiatives in under-represented regions, but is still very sparse in Africa and nonexistent in Latin America. The main reasons behind this lack of success are essentially linked to long-term sustainability and funding of the social infrastructures.

Other examples exist too: the World Meteorological Organization Information System (WIS, WMO, 2014) and the Group on Earth Observations GEOSS also contribute to the same vision and represent major initiatives to enhance international coordination in order to provide the basis for common infrastructures. More recently, the Research Data Alliance, an action-orientated international framework currently supported by national science funders in Australia, Europe, and the United States, was established to help overcome technical and social barriers hampering data sharing and reuse.

The challenges facing society are multidisciplinary by nature, and therefore global data-related efforts such as the ones mentioned need contributions from all domain- and discipline-specific data communities, including polar data. We will concentrate on at least two aspects of involvement and contribution in the following two sections: the involvement of key stakeholders and the promotion of good practices.

3.1 Involving the stakeholders

The Antarctic community has an existing and long-standing international data management effort operating under the umbrella of ICSU's Scientific Committee for Antarctic Research and the Antarctic Treaty (Finney, 2013). The Arctic polar data community is also increasingly concerned with data preservation and sharing, and efforts have started under the auspices of the International Arctic Science Committee (IASC), an ICSU

Associate Member, and the Arctic Council to increase awareness about data issues (IASC, 2013). These initiatives bringing together national, regional, and international data repositories and data service providers to coordinate their efforts have various levels of maturity and success but are essential parts of the global infrastructure needed to ensure open access and long-term preservation of essential polar data to the benefit of the international research community. Additional efforts are needed to better coordinate and work with other key stakeholders, such as libraries, science funders, and publishers to maximise the benefits of existing national investments and global initiates such as ICSU-WDS, WIS, RDA, and others.

3.2 Promoting good practices

One of the key roles international data-related initiatives play is to promote good practices amongst communities in order to improve the overall performance of data systems, better respond to requirements of science funders and policy makers, and ultimately benefit scientific research. These good practices include the implementation of open data policies, the development of trusted systems and long-term funding strategies to support data repositories, and endorsement of change in scientific practices to require sharing and citing data.

Open data policies and good practices in data management were adopted but not necessarily fully implemented during IPY. However, they paved the way to and influenced policies currently in place at the global level, such as the GEOSS Data Sharing Principles (Group on Earth Observations, 2008) and the newly developed IASC Data Policy (IASC, 2013). A wider diffusion and better implementation of such policies and practices in the scientific research community is needed and can be facilitated by adapting these to specific disciplinary requirements where appropriate. For example, the concept of 'Ethically Open Access' is articulated in the IASC Statement of Principles and Practices for Arctic Data Management to reconcile the requirements for openness and the legitimate requirements to protect privacy of human subjects, traditional knowledge, and conservation of species.

Publishing data, including the use of permanent identifiers such as Digital Object Identifiers, has also gained a lot of international traction. Mechanisms for publishing and citing data are promoted and used by some of the leading polar data management services but are not widely accepted in the developing polar data management networks. Several international efforts to establish the publishing and citing of data as accepted norms in the scholarly world are currently underway. In the area of data citation, for example, long-standing international efforts have recently culminated with a coalition of organizations working in this area, the *Data Citation Synthesis Group*, to achieve international agreement on *Data Citation Principles* to be widely recognized, endorsed, and implemented in academia (Data Citation Synthesis Group, 2014). Similarly, ongoing international initiatives such as the Publishing Data Working Groups coordinated by ICSU-WDS and RDA are bringing together various stakeholders, data centres, data service providers, publishers, funders, and bibliometrics providers, to establish an international framework for publishing data (WDS Data Publication WG, 2014). Publishing and citing data are good practices, offering incentives to data practitioners, in the form of scientific publications and citations, and benefits to the scientific community by improving accessibility and usability of datasets.

Certification of data repositories is another mechanism to promote good practises and improve trust in data infrastructure. A number of synergetic certification procedures co-exist, ranging from the rigorous International Organization for Standardization certifications to more community-based norms such as the ICSU-WDS and Data Seal of Approval accreditations. The organizations behind these two norms are currently exploring ways to harmonize their catalogues of criteria to offer a framework for baseline certification covering Natural and Social Sciences. Many of the challenges posed by IPY in terms of data management could have been easier to solve if a network of certified polar data repositories was available to respond to the needs of the various research projects involved. For this reason, the polar data community needs to adopt certification procedures proactively for its relevant data repositories and data services to align their capacities with those similar in other domains and thus ensure proper integration of polar data in the global scientific endeavour.

4 CONCLUSION

IPY advanced polar data stewardship and improved data availability and data science practice. To continue to address the complexities of diverse data, the community needs to grow, constantly improve its practices, and build relationships globally and locally within disciplines and regions. Periodic conferences, such as the recent International Forum on 'Polar Data Activities in Global Data Systems' in Tokyo, and assessments of the state of polar data practice should continue under the umbrella of the relevant national, regional, and international polar organizations and in collaboration with international research and data-related initiatives. It is important that the polar data community strengthen itself, but it also must reach out beyond that community to build relationships and share knowledge with broader global organizations.

So far, the weaknesses and relative lack of coordination in global scientific data systems are hindering the full realization of societal benefits expected from taxpayer-funded research. The reasons behind this slow progress are diverse, and range from relatively easy to solve technical issues, such as metadata formats, to the more difficult to tackle sociopolitical obstacles to transnational harmonization. Another important barrier is insufficient recognition for data management practitioners' work in the scientific community on the one hand and the dearth of new and sustainable funding mechanisms to support internationally coordinated data eInfrastructure needed by the scientific community on the other hand. Data science must be considered an integral part of science in general. It must be included in the training of the next generation of scientists and be funded as part of their scientific activities.

Much remains to be solved to build an internationally coordinated research data infrastructure that provides openly accessible and usable scientific data. In the past, initiatives such as the ICSU World Data Centres demonstrated how a flexible international coordination mechanism, based solely on national capacities, could deliver successful long-term data preservation and accessibility for a specific domain of research. However, today's scientific endeavour, the societal challenges we face, the amount of funding available, and the volumes of data produced have dramatically changed. This new landscape requires innovative, adaptive solutions to accommodate and achieve flexible collaboration and coordination between domain-specific data communities, such as the polar-related research community, enabling them to advance their own activities and at the same time to open to and link with other domains.

5 REFERENCES

Aronova, E., Baker, K.S., & Oreskes, N. (2010) Big Science and Big Data in Biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) Network, 1957–Present. *Historical Studies in the Natural Sciences 40*(2), pp 183–224.

Baeseman, J. & Pope, A. (2011) APECS: Nurturing a new generation of polar researchers. *Oceanography 24*(3), p 219. Retrieved September 28, 2014 from the World Wide Web: http://dx.doi.org/10.5670/oceanog.2011.73

Battrick, B. (Ed.) (2005) *Global Earth Observation System of Systems GEOSS; 10-Year Implementation Plan Reference Document,* Noordwijk: ESA Publications Office.

Carlson, D.J. (2010) Why do we have a 4th IPY? In Barr, S. & Lüdecke, C., (Eds), *The History of the International Polar Years (IPYs)*, Berlin: Springer-Verlag.

Carlson D.J. (2011) A lesson in sharing. *Nature 469*, p 293.

Data Citation Synthesis Group (2014) Joint Declaration of Data Citation Principles – FINAL. Retrieved March 02, 2014 from the World Wide Web: http://www.force11.org/datacitation

Future Earth Transition Team (2014) Future Earth Initial Design, p 41 and Annex 4. Retrieved February 17, 2014 from the World Wide Web: http://www.icsu.org/future-earth/media-centre/relevant_publications/future-earth-initial-design-report

Group on Earth Observations (2008) The GEOSS Data Sharing Principles. Retrieved February 17, 2014 from the World Wide Web: https://www.earthobservations.org/documents/geo_vii/07_GEOSS%20Data%20Sharing%20Action%20Plan%20Rev2.pdf

G8+O5 Global Research Infrastructure Sub Group on Data (2011) Draft Report. Retrieved February 17, 2014 from the World Wide Web: http://cordis.europa.eu/fp7/ict/e- infrastructure/docs/g8.pdf

Harris, R. (2012) ICSU and the Challenges of Big Data in Science. *Research Trends 30* (Section 4), pp 11.

Kim Finney (2013) *SCAR Report: Data and Information Management Strategy (DIMS)*. Retrieved February 17, 2014 from the World Wide Web: http://scadm.scar.org/0files/SCAR_DIMS_34.pdf

IASC (2013) *Statement of Principles and Practices for Arctic Data Management*. Retrieved February 17, 2014 from the World Wide Web: http://www.iasc.info/home/iasc/data

Parsons, M., Godøy, Ø., LeDrew, E., de Bruin, T., Danis, B., Tomlinson, S., & Carlson, D. A. (2011) Conceptual framework for managing very diverse data for complex interdisciplinary science. *J. of Information Science 37*(6), pp 555–569.

Parsons, M., de Bruin, T., Tomlinson, S., Campbell, H., Godøy, Ø. & LeClert, J. (2011) The state of polar data—the IPY experience. In Krupnik, I., Allison, I., Bell, R., Cutler, P., Hik, D., López-Martínez, J., Rachold, V., Sarukhanian, E., & Summerhayes, C., (Eds.), *Understanding Earth's Polar Challenges: IPY 2007–2008*, Edmonton: CCI Press.

Rapley, C., Bell, R., Allison, I., Bindschadler, R., Casassa, G., Chown, S., Duhaime, G., Kotlyakov, V., Kuhn, M., Orheim, O., Pandey, P.C., Petersen, H.K., Schalke, H., Janoschek, W., Sarukhanian, E., & Zhang, Z. (2004) *A Framework for the International Polar Year, 2007–2008,* Paris: International Council for Science.

WDS Data Publication WG (2014) Data Publication Working Group. Retrieved March 02, 2014 from the World Wide Web: https://www.icsu-wds.org/community/working-groups/data-publication

WMO (2014) WMO Information System. Retrieved February 17, 2014 from the World Wide Web: http://www.wmo.int/pages/prog/www/WIS/

(Article history:Available online 17 October 2014)