Managing very diverse data for complex, interdisciplinary science

Mark Parsons, Øystein Godøy, Ellsworth LeDrew, Taco de Bruin, Bruno Danis, Scott Tomlinson, David Carlson

parsonsm@nsidc.org

The challenge of massive data volumes receives much more attention than the challenges of data diversity in modern data-intensive science. We use the experience of the International Polar Year (IPY) to examine data management approaches that address issues around complex *interdisciplinary* science. We find that while technology is a critical factor in addressing the interdisciplinary dimension of e-science, the technologies developing for exa-scale data volumes are not the same as what is needed for extremely distributed and heterogeneous data. A much simpler and flexible approach is needed. More importantly, there is a need for both technical and cultural adaptation. We take a holistic, science and technology studies approach that lead us to suggest several short and long-term strategies to facilitate a socio-technical evolution in the overall science data ecosystem.

> We are persuaded by Latour that *the important questions concern the flow of objects and concepts through the network of participating allies and social worlds*. (Star and Griesemer 1989. p. 389]

Our vision is that data should be:

Discoverable

Users need not only to search but also to explore. And they need specialist guidance along the way.

Not a "one stop shop"

Rather than a one-stop shop, a better metaphor is 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.

But a Grand Bazaar!

photo by Frank Kovalchek (CC-BY)

Useful

Useful (and usable) is in the eye of the beholder, requiring flexibility by data managers and data providers.

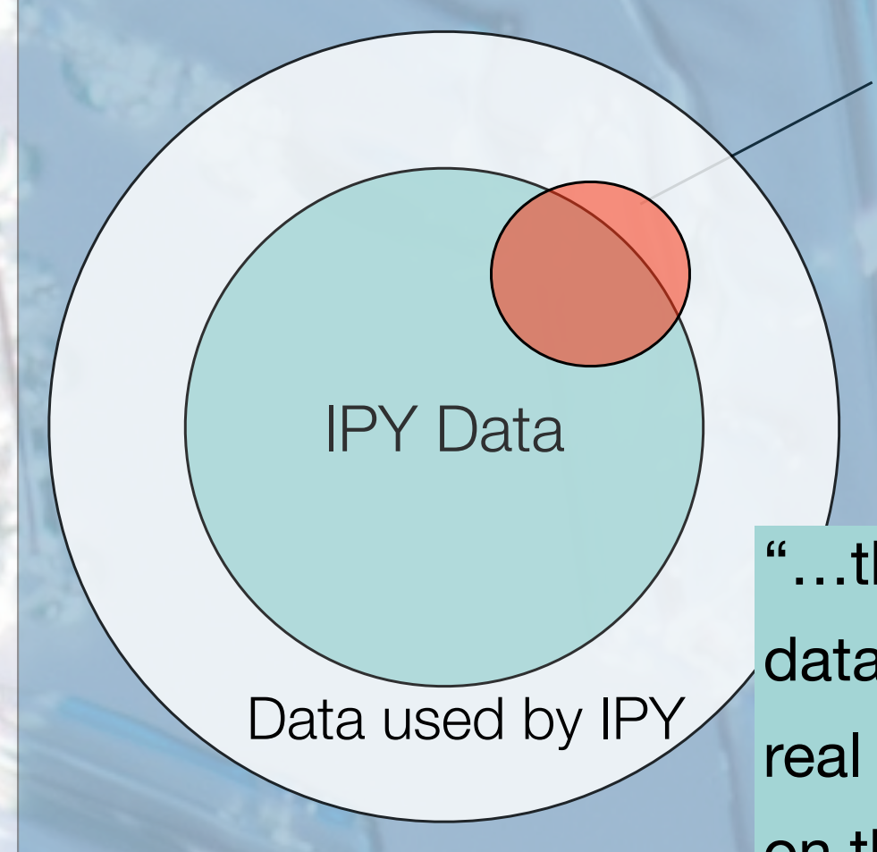
With such a diversity of users, needs, and applications, much more than just the data are necessary. Users need the data to be coherent in form and semantics with their models and analysis tools. They need rich documentation fully describing data uncertainties and fitness for use. They need context and background about algorithms, calibrations, and methods. Furthermore, the decision makers do not really need data at all but rather information presented in compelling, readily interpretable ways. Effective data and information display can encourage greater data sharing, increase understanding of complex processes, and enable wiser decisions.

This means there will be an increasing need for informatics specialists in the twenty-first century. These specialists will need to include not only computer scientists and systems engineers grappling with complex technical issues, but also data scientists, data curators, librarians, data 'wranglers', information designers, and even artists grappling with social systems and improving human understanding. Society will increasingly rely on those professionals who act as translators to make complex, distributed data accessible and useful.

Open

A forward-looking, ethics-based data policy moving toward a new information commons.

IPY Data Policy



Special Cases:
•Human subjects
•Intellectual property of LTK
•Where data release may cause harm

"…the IPY Joint Committee requires that IPY data, including operational data delivered in real time, are made available fully, freely and on the shortest feasible timescale."

The Polar Information Commons

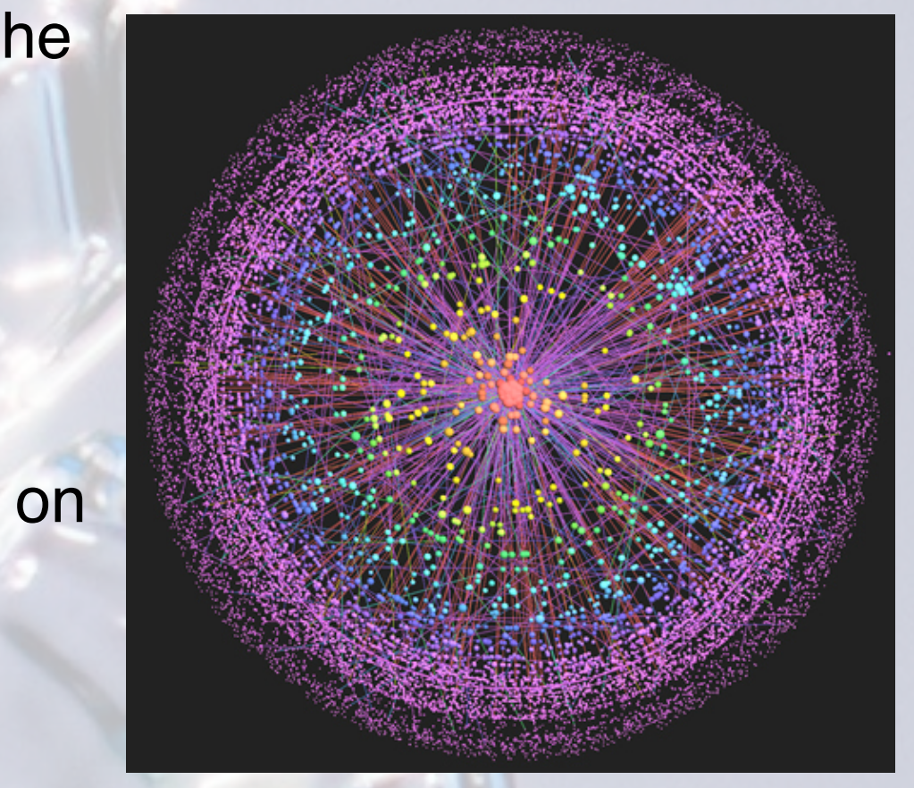
The core principle of the PIC is that data should not be seen as the property of an individual, but should instead be seen as a public good. And like any public good it should be used and shared responsibly and ethically.

This ethical behavior cannot be mandated but instead should be based on a set of community norms. Initial norms include concepts like attribution of data providers, adequate description of data quality and uncertainty, and sharing derivative works. http://polarcommons.org

Linked

Data are more relevant and useful if they are associated and explicitly linked with other data, especially when they are linked in a way that computers can readily interpret.

Most Earth system science data are managed in hierarchical file systems and relational databases. Connections to other data are through structured descriptions of hierarchies in metadata, XML schemas, and file structures or through primary key relationships in databases. A more specific concept of linked data is through the Semantic Web. Earth science has relied on metadata catalogs not only for data discovery but also to associate related data through hierarchies, structured relationships, and defined keywords or vocabularies. While these registries describe the data well, they do not provide consistent access to the data. There may be a direct link to the data, but too often it is just a link to another web site, or there is no link at all. So while metadata is increasingly linked across catalogs, the actual data remain disconnected. Furthermore, metadata catalogs do not often capture the meaning, or the semantics, of the actual content of the metadata.



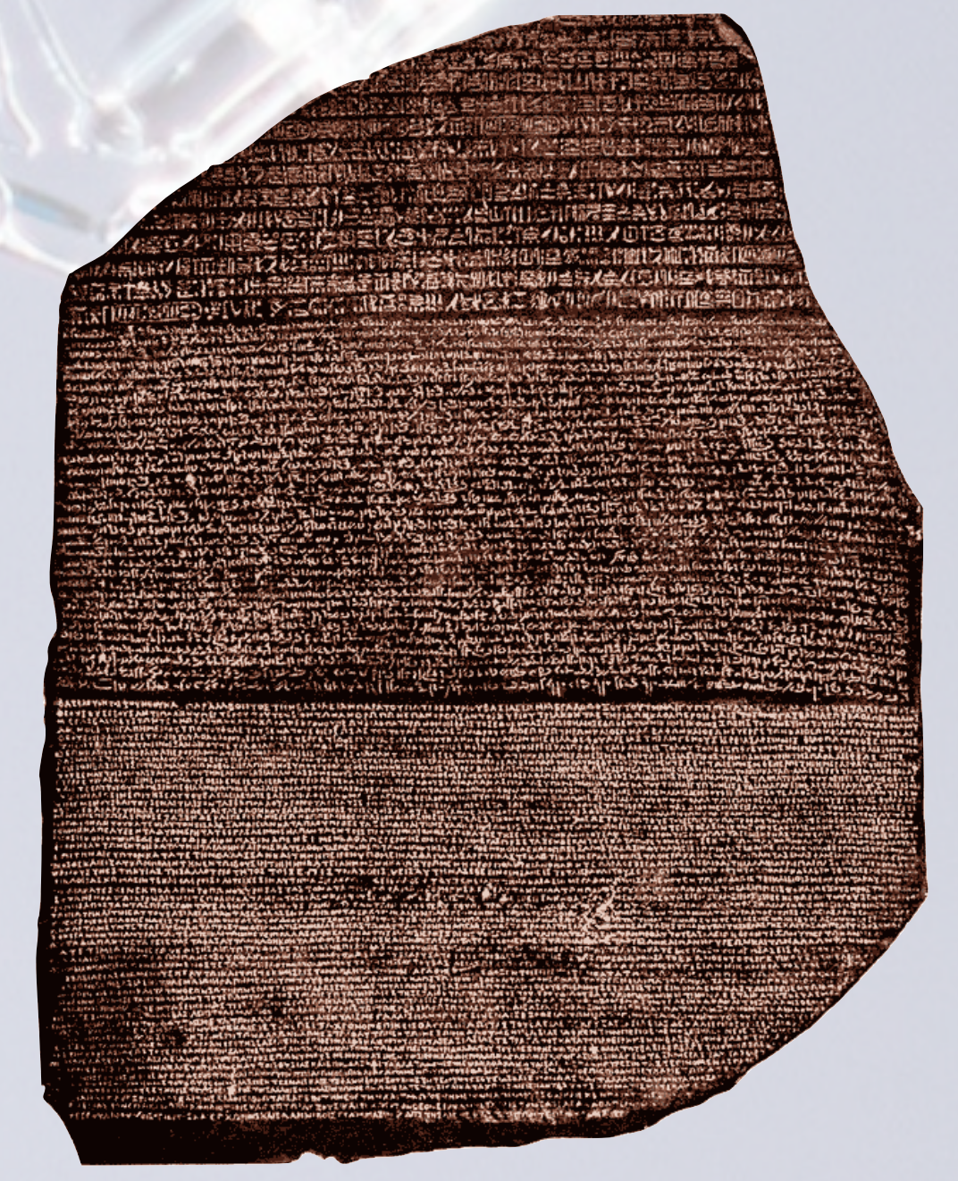
A semantically rich, easily extensible, linked-data approach seems most suitable for interconnecting diverse research data across existing disciplinary data silos. It allows adaptive and iterative development, without the construction of complex hierarchical structures or heavyweight standards at the outset. But linked data is an evolving approach that still requires substantial agreement on detailed standards, such as ontologies, to realize its full potential.

Safe

Safe from hackers, from obsolescence, from undocumented change, from loss, and from the ravages of time.

Keeping data safe, now and for the long term, is the best-understood but most difficult challenge of the fourth paradigm. By safe, we simply mean that data integrity is recorded and preserved and the data remain usable for future generations. Multiple, high-level studies have highlighted the critical need for and challenges of data security and long-term preservation. There is even a well-regarded international standard on what an 'Open Archive Information System' needs to do. Unfortunately, despite this broad understanding, most research data are likely to be lost. IPY starkly revealed this disparity between theory and practice around the world. While it is yet early for a full assessment, at the completion of IPY it appeared that only 30 of 124 large IPY science projects (24%) had adequately planned for long-term preservation.

IPY has made an impact, though. More funding agencies consider data sharing a requirement for continued support, and new archives are being established in several countries to preserve IPY data. The new ICSU World Data System (WDS) has taken on the preservation of IPY data as a major priority. The WDS promises to be a reification of the data ecosystem, but it will take time to grow and requires active involvement of major ecosystem components, notably sustained institutions and funding agencies.



Reflections

- Data can act as *boundary objects* that help communicate across disciplines.
- Participation in the minutiae of data creation can be a means to help form *scientific identity and community*. Data managers need to participate in data creation because it helps engender trust and collaboration, but it also makes the boundary object, the data, more robust.
- While, it is important to consider user needs and perspectives, it is equally important to *consider data creator perspectives* to effectively capture contextual knowledge.
- "*The subject knowledge view of relevance is fundamental to all other views of relevance,* because subject knowledge is fundamental to communication of knowledge." (Saracevic, 1975, his emphasis)
- There is often complete separation between the data creator and their ultimate steward (if there is one). We lack a *"keystone species"* in the ecosystem—the data scientist, the mediator, the translator.
- Data need to flow smoothly through the ecosystem.
- Scientists must see themselves as part of the data ecosystem, and this will require flexible tools and the time, patience, and outreach of data managers.
- We need to create and link "complex e-science objects" and capture tacit knowledge
- "We must ask users to meet designers halfway by learning their language and developing an understanding of the design domain. If designers are at fault for assuming that all user requirements can be formally captured and codified, users are often equally at fault for expecting 'magic bullets'—technical systems that will solve social or organizational problems." —Star and Ruhleder (1996)
- Sponsors are part of the ecosystem too.

Facilitating a socio-technical evolution

Technical Strategies

Short term:

- Develop open, cloud-based approaches of data broadcasting and customized aggregation.
- Make data and metadata available through a multiple protocols and formats to serve various communities.
- Data systems need to start simple and iterate to expand their interconnection with other systems and user communities.
- Developers need to work closely with data providers to improve acceptance and use of standards.

Long term:

- Informatics research needs to explore ways to better define, describe, automatically create, and interrelate complex e-science objects across disciplines and data systems.

Cultural Strategies

Short term:

- Funding agencies must require data management plans as part of basic research proposals and then fund archives to support the plans.
- Basic data management needs to be included in the core scientific curriculum.
- Data providers should receive formal recognition through data citation.
- Data managers need to establish close working relationships with their data providers as well as their users, based on mutual trust.

Long term:

- Data scientists need to continue to professionalize their discipline
- When we consider data a common good, this suggests that preservation of data should be a broad societal cost.

This poster is derived from:
Parsons, MA, Ø Godøy, E LeDrew, TF de Bruin, B Danis, S Tomlinson, and D Carlson. 2011. A conceptual framework for managing very diverse data for complex interdisciplinary science. *Journal of Information Science*. http://dx.doi.org/10.1177/0165551511412705

World Climate Research Programme Open Science Conference, 25 October 2011