BUILDING ON THE INTERNATIONAL POLAR YEAR: DISCOVERING INTERDISCIPLINARY DATA THROUGH FEDERATED SEARCH

L Yarmey[1]* and S J Khalsa[1]

[1]National Snow and Ice Data Center, University of Colorado Boulder, Boulder, CO 80309, USA
**Email:* lynn.yarmey@nsidc.org

ABSTRACT

The legacy of the International Polar Year 2007–2008 (IPY) includes advances in open data and meaningful progress towards interoperability of data, systems, and standards. Enabled by metadata brokering technologies and by the growing adoption of international metadata standards, federated data search welcomes diversity in Arctic data and recognizes the value of expertise in community data repositories. Federated search enables specialized data holdings to be discovered by broader audiences and complements the role of metadata registries such as the Global Change Master Directory, providing interoperability across the Arctic web-of-repositories.

Keywords: Information infrastructure, Metadata and systems interoperability, Federated data search, Metadata brokering, Interdisciplinary data discovery

1 INTRODUCTION

Modern polar research benefits from the complex history of Arctic and Antarctic science. Observations over time and across geospatial scales and domains are crucial for setting baselines and for understanding the rapid changes in these key regions. The International Polar Year 2007–2008 (IPY) played a critical role in promoting data sharing and in offering central coordination for observations. The IPY Data and Information Service (IPYDIS) identified early the unique needs of interdisciplinary, international research and introduced the 'rigorous yet collaborative' approach of a *union catalogue* (Parsons, 2006). This catalogue would allow for disparate search strategies and interfaces to access IPY data and resources. The experiences from IPYDIS union catalogue design and implementation are well documented (Parsons, Godøy, LeDrew, de Bruin, Danis, Tomlinson, et al., 2011). Community discussion on these important topics has continued at the IPY 2012 workshop of the Arctic Data Coordination Network (2012).

Despite these significant steps, polar data discovery and access challenges remain. Many data repositories, funded by different agencies, nations, and operational and industry groups maintain separate catalogues and systems designed to meet different needs. Given this complex and diverse landscape of resources, researchers and others looking to reuse data need a good deal of insider knowledge, time, and luck to discover, access, and understand data. Leveraging IPY contributions as well as recent technical advances, metadata brokering addresses data discovery challenges by mediating across distributed systems. Metadata brokering tools, the focus of an EarthCube Building Blocks BCube project funded by the National Science Foundation (NSF), work with different access mechanisms, update schedules, and metadata standards. One type of brokering configuration regularly aggregates heterogeneous metadata into a single system. Federated search portals such as the Advanced Cooperative Arctic Data and Information Service (ACADIS) Arctic Data Explorer then leverage the brokered metadata to facilitate searches across many distributed sources simultaneously. Federated search presents an opportunity to advance polar cyberinfrastructure towards the vision of a web-of-repositories (Baker & Yarmey, 2009) through coordination of standards, infrastructure, and resources to support critical polar science.

2 DISCUSSION

Federated data search honours the rich legacy of polar research by enabling diversity in participating repositories. Polar data repositories and services often form in response to community-specific needs. For example, monitoring stations have different data requirements than seasonal biological surveys or one-time soil

moisture projects. Communities have different metadata and data content standards and encodings, vocabularies, and access protocols. Metadata brokering enables all of this diversity to exist while minimizing the amount of additional standardization needed, reducing the burden placed on repository managers. Repositories interested in participating in global and polar cyberinfrastructure efforts are not required to send data to a central system or manage their data in a prescribed manner. Federated data search through metadata brokering bridges the distributed legacy of polar science.

2.1 Technology—Metadata brokering

ACADIS is using the brokering technologies developed by the Italian Centre for National Research – Earth and Space Science Informatics Laboratory (ESSI-Lab, 2014). Their Brokering Framework includes discovery, access, semantic, workflow, and quality brokers. The discovery broker component, called GI-cat (Nativi & Bigagli, 2009), was utilized in the work we report on here.

GI-cat can access metadata through a variety of exchange protocols including the Open Archives Initiative – Protocol for Metadata Harvesting, Thematic Real-time Environmental Distributed Data Services, and many others. Once metadata are accessed and harvested, GI-cat aggregates the records into a central database. Alternative brokering models and tools distribute queries 'on the fly' and aggregate the results for presentation to the user. In either case, metadata are translated from native standards into a common schema, the International Standards Organization (ISO) 19115 in the case of GI-cat, using crosswalks. Once harmonized and indexed, a search query sent through web services (e.g., OpenSearch) produces a results set that is displayed based on a defined relevance ranking algorithm. Resource links in the metadata enable users to view the original metadata record for a dataset of interest, with the potential to build additional functionality, such as download or transformation, based on web services. Figure 1(a) shows the Arctic Data Explorer's high-level architecture, highlighting the repositories, metadata feeds, GI-cat harvest and broker layer, SOLR query handling, web services, and web portal components.

GI-cat along with the access and semantic components of ESSI-Lab's Brokering Framework comprise the Discovery and Access Broker (DAB) used by the Global Earth Observation System of Systems. The DAB has proven to be a viable mechanism for aggregating metadata on a scale even larger than IPY.

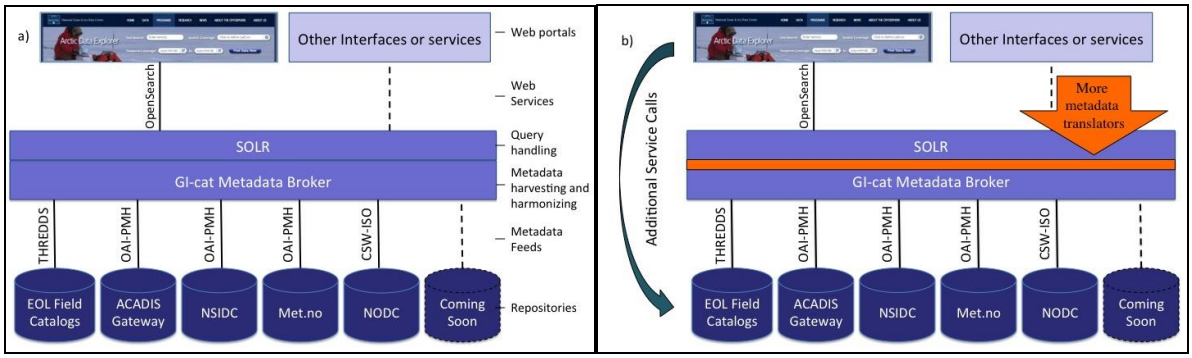

Figure 1. (a) Flexible, extensible architecture of the Arctic Data Explorer metadata brokering stack (available at http://nsidc.org/acadis/search/). Multiple repositories are included in the search and many more are planned. (b) Core Arctic Data Explorer architecture, along with the secondary metadata translation layer and the planned Additional Service Calls to applicable originating repositories

2.2 Experience with federated data search

The ACADIS Arctic Data Explorer experience has shown that to have a successful, sustainable, open source metadata brokering product, a few points should be considered. Most important is recognizing that the biggest challenges in metadata brokering are not necessarily technical. For example, relationships among developers across participating institutions are key, and consistent, honest, and timely communication with stakeholders is required. The following stakeholders should be included: scientists providing data, scientists searching for data, developers, data curators, system architects, technical operations staff, project managers, metadata experts, web usability experts, other federated data search efforts, and funders. Ongoing coordination and alignment are important to success in metadata brokering. In addition to working closely with stakeholders, it is vital to

identify the primary audience for the brokered application and proactively seek usability feedback from them early and often (Yarmey & Wilcox, 2012).

2.3 Challenges and next steps

Significant progress has been made towards comprehensive federated data search through metadata brokering though challenges remain. Experiences thus far have highlighted additional technical, social, and sociotechnical elements necessary to achieve scientific goals. Examples include: long-term maintenance and resourcing system scaling, lack of governance structures, meaningful relevance ranking of thousands of search results to help searchers find what they need, building trust into systems, and others.

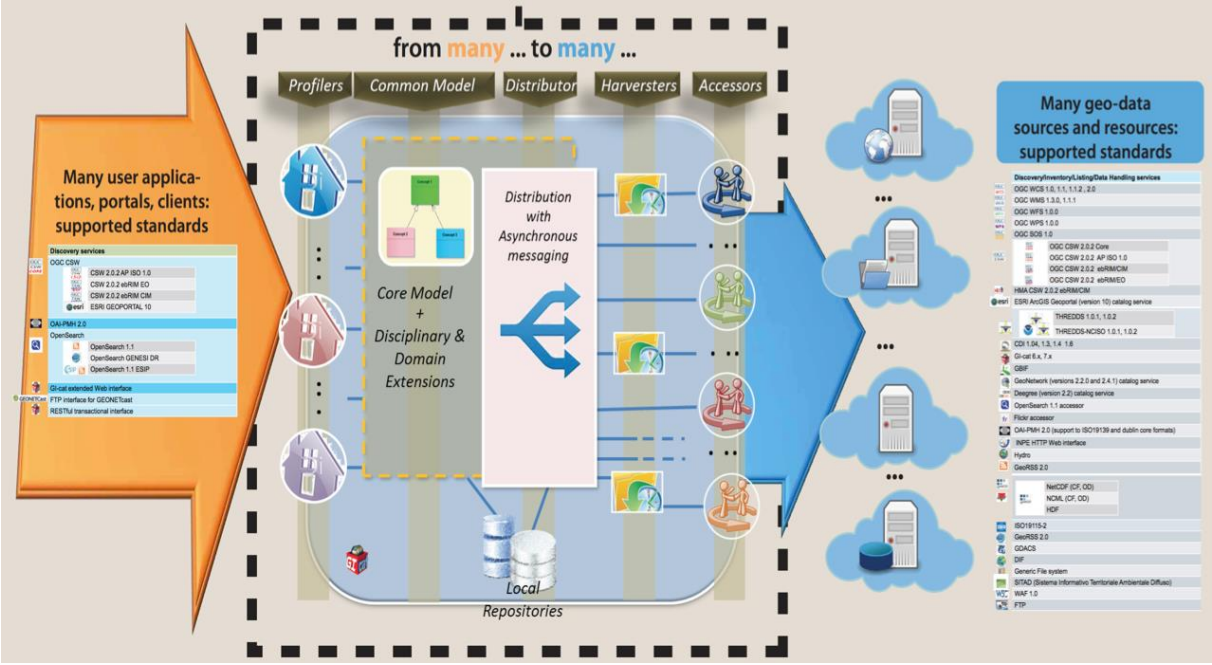

Figure 2. Planned architecture of EarthCube Building Blocks BCube project (Khalsa, Pearlman (J); Nativi, Pearlman (F); Parsons, Browdy, et al., 2013)

Inconsistent application of metadata standards presents a core challenge to distributed discovery. In an ideal situation, metadata structured into a standard such as ISO 19115 could be translated to other standards based on a single crosswalk application. The metadata in standard form would be semantically, syntactically, and structurally interoperable with other metadata in that same standard form. However, while many communities have chosen and enacted metadata standards, the content encoded in these standards has rarely proven to be actually standardized. Experiences with the Arctic Data Explorer thus far have shown problems such as inconsistent content in the same standard field, similar content in different standard fields, and other barriers to the straightforward use of existing crosswalks. For example, a 'data provider' metadata field in a standard might be applied by two different organizations to contain either the originating researcher name(s) or the name of the data centre now serving that data. A metadata broker will see the standard and will apply the same crosswalk, equating the two entries inappropriately. In the long term, semantic technologies may help address content reconciliation though short-term approaches are also needed. The Arctic Data Explorer has a secondary layer of additional metadata mappings on top of the initial crosswalk application to ensure queries access truly standardized content (Figure 1(b)). Such remapping moves content between fields so as to normalize across different providers. More work will be needed to ensure comprehensive interoperability of metadata and metadata standards as brokering solutions scale across more diverse repositories. Short-term next steps include community negotiated and enacted best practices, such as guidance on what content belong in common fields of different metadata standards. In the long-term, increased consideration should be given to the participatory models for development of metadata standards (Yarmey & Baker, 2013).

Many of the above issues will be explored and addressed through the recently funded 'BCube' project, a component of the United States NSF's EarthCube program (Figure 2). EarthCube aims to guide the development of a cyberinfrastructure to support multidisciplinary collaboration in the geosciences. BCube will research the social and technical aspects of brokering in support of science in the polar, oceans, hydrology, and

weather/climate domains. One facet of this research will be to explore how different instances of brokers can interact and share information about the resources that each broker mediates.

3 CONCLUSION

Federated data search, made possible by metadata brokering technologies, begins to address the problem of finding data of interest in a myriad of diverse, isolated repositories. The experience of the ACADIS Arctic Data Explorer in doing federated data search in the Arctic is informing trans-polar data discovery and access efforts. Ongoing communication, coordination, research, and development are needed to address challenges.

4 ACKNOWLEDGEMENTS

ACADIS, funded under the United States National Science Foundation Award ARC 1016048, partners the University Corporation for Atmospheric Research with the National Center for Atmospheric Research and the National Snow and Ice Data Center. The BCube project is funded under EarthCube Award ICER 1343802 by the National Science Foundation. The authors would like to thank the ESSI-Lab and NSIDC development teams.

5 REFERENCES

Arctic Data Coordination Network (2012) Arctic Data Coordination Network (ADCN) Workshop Report. *IPY 2012*, Montréal, Quebéc, Canada.

Baker, K.S. & Yarmey, L. (2009) Data Stewardship: Environmental Data Curation and a Web-of-Repositories. *International Journal of Digital Curation, 2*(4), pp 12–27.

Italian Centre for National Research – Earth and Space Science Informatics Laboratory (2014) Retrieved May 28, 2014 from the World Wide Web: http://essi-lab.eu

Khalsa S.J., Pearlman, J., Nativi, S., Pearlman, F., Parsons, M., Browdy, S., & Duerr, R. (2013) *Brokering for EarthCube Communities: A Road Map*. Retrieved May 28, 2014 from the World Wide Web: http://dx.doi.org/10.7265/N59C6VBC

Nativi, S. & Bagagli, L. (2009) Discovery, Mediation, and Access Services for Earth Observation Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 2*(4), pp 233–240.

Parsons, M. A. (2006) International Polar Year Data Management Workshop, 3–4 March 2006. *Glaciological Data Series*, GD-33.

Parsons, M. A., Godøy, Ø., LeDrew, E., de Bruin, T. F., Danis, B., Tomlinson, S., & Carlson, D. (2011) A Conceptual Framework for Managing Very Diverse Data for Complex Interdisciplinary Science. *Journal of Information Science 37*(6), pp 555–569.

Yarmey, L. & Baker, K.S. (2013) Towards Standardization: A Participatory Framework for Scientific Standard-Making. *International Journal of Digital Curation 8*(1), pp 157–172.

Yarmey, L. & Wilcox, H. (2012) Brokering technologies as a framework for collaborative data curation. *American Geophysical Union Fall Meeting*, San Francisco, California, USA.

(Article history:Available online 17 October 2014)