

RESEARCH PAPER

Data and Metadata Brokering – Theory and Practice from the BCube Project

Siri Jodha Singh Khalsa

National Snow and Ice Data Center, Cooperative Institute for Research in Environmental Science, University of Colorado, Boulder, Colorado, USA

sjsk@nsidc.org

EarthCube is a U.S. National Science Foundation initiative that aims to create a cyberinfrastructure (CI) for all the geosciences. An initial set of “building blocks” was funded to develop potential components of that CI. The Brokering Building Block (BCube) created a brokering framework to demonstrate cross-disciplinary data access based on a set of use cases developed by scientists from the domains of hydrology, oceanography, polar science and climate/weather. While some successes were achieved, considerable challenges were encountered. We present a synopsis of the processes and outcomes of the BCube experiment.

Keywords: interoperability; brokering; middleware; EarthCube; cross-domain; socio-technical

Genesis and Objectives of EarthCube

In 2011 the U.S. National Science Foundation initiated EarthCube, a joint effort of NSF’s Office of Cyberinfrastructure (OCI), whose interest was in computational and data-rich science and engineering, and the Geosciences Directorate (GEO), whose interest was in understanding and forecasting the behavior of a complex and evolving Earth system. The goal in creating EarthCube was to create a sustainable, community-based and open cyberinfrastructure for all researchers and educators across the geosciences.

The NSF recognized that currently there was no infrastructure that could manage and provide access to all geosciences data in an open, transparent and inclusive manner, and that progress in geosciences would be increasingly reliant on interdisciplinary activities. Therefore, a system that enabled the sharing, interoperability and re-use of data needed to be created.

Similar efforts to provide the infrastructure needed to support scientific research and innovation is underway in other countries, most notably in the European Union guided by the European Strategy Forum on Research Infrastructures (ESFRI) and in Australia under the National Collaborative Research Infrastructure Strategy (NCRIS). The goal of all these efforts is to provide scientists, policy makers and the public with computing resources, analytic tools and educational material, all within an open, interconnected and collaborative environment.

The Nature of Infrastructure Development

The building of infrastructure is as much a social endeavor as technical one. Bowker, et al. (2010) emphasized that information infrastructures are more than the data, tools and networks comprising the technical elements, but also involve the people, practices, and institutions that lead to the creation, adoption and evolution of the underlying technology. The NSF realized that a cyberinfrastructure, to be successful, must have substantial involvement of the target community through all phases of its development, from inception to deployment. In fact, studies have shown infrastructure evolves from independent and isolated efforts and there is not a clear point where “deployment” is complete (Star and Ruhleder, 1996). The fundamental challenge was the heterogeneity of scientific disciplines and technologies that needed to cooperate to accomplish this goal, and the necessity of getting all stakeholders to cooperate in its development. A compounding factor is that while technology evolves rapidly, people’s habits, work practices, cultural attitudes towards

data sharing, and willingness to use other's data, all evolve more slowly. How the relationship of people to the infrastructure evolves determines whether it succeeds or fails.

A significant element of NSF's strategy for building EarthCube was to make it a collective effort of geoscientists and technologists from the start, in hopes of ensuring that what was developed did indeed serve the needs of geoscientists and would in fact find widespread uptake. A series of community events and end-user workshops spanning the geoscience disciplines were undertaken with the dual goals of gathering requirements for EarthCube and building a community of geoscientists willing to engage with and take ownership of the EarthCube process.

NSF began issuing small awards to explore concepts for EarthCube. These were followed by the funding of an initial set of "building blocks" meant to demonstrate potential components of EarthCube. The Brokering Building Block (BCube) was one of these awards. BCube sought both to solve real problems of interoperability that geoscientist face in carrying out research, while also studying the social aspects of technology adoption.

The Challenge of Cross-Disciplinary Interoperability

Interoperability has many facets and can be viewed from either the perspective of systems or people. Systems are interoperable when they can exchange information without having to know the details of each other's internal workings. Likewise, people view systems or data as interoperable when they don't have to learn the intricacies of each in order to use them. When systems are interoperable users of those systems should have uniform access and receive harmonized services and data from them. This is the vision of EarthCube. Delivering on that vision can be considered the 'Grand Challenge' of information technology as applied to the geosciences.

The reason that achieving interoperability across the geosciences is so challenging is that the many scientific fields that comprise the geosciences all have their own methods, standards and conventions for managing and sharing data. The sophistication of the information technologies that have been adopted in each community, the degree of standardization on data exchange formats and vocabularies, the amount of centralization in data cataloguing, and the openness to sharing data all vary greatly.

The methods of achieving interoperability across distributed systems can be categorized as shown in **Table 1**.

Since disciplines will always use different standards for encoding, accessing and describing data, the first option is not a realistic one for the geosciences. The second method is currently in wide use within the geosciences, such as GBIF (Edwards, Lane and Nielsen, 2000), which harvests metadata from multiple external systems and then maps the metadata, which are served through different protocols and use different schemas, to a common standard. Systems such as ERDAAP (Simons and Mendelssohn, 2012; Delaney, Alessandrini and Greidanus, 2016) act as servers accessing disparate datasets and serving them through a common interface. What BCube explored was the possibility that a broker, mediating the interactions between many systems serving data and many systems requesting data, could be established as a shared service, i.e. as infrastructure, without being tied to any particular repository or user portal.

Edwards et al. (2007) show that technical infrastructures such as electrical grids and railroads evolve in stages, and the final stage is "a process of consolidation characterized by *gateways* that allow dissimilar systems to be linked into networks". Brokering is such a gateway, applied in the context of information systems. While brokering technologies such as CORBA¹ have been in existence since the 1990's, their application typically requires participants in a network to install software packages that enable interfacing

Method	Requirements	Benefits
Adherence to common standards	Uniformity in system configuration	De facto interoperability
Gateways and translators	Installation and maintenance of custom or 3 rd party software	Can adapt to new or changing protocols and standards
Brokers as infrastructure, 3rd party mediation	Creation and maintenance of brokering framework with custom adapters	Provides 2-way translations between disparate systems Removes burdens of interoperability from data provider

Table 1: Methods for achieving interoperability.

¹ The Common Object Request Broker Architecture (CORBA), a standard defined by the Object Management Group (OMG), is designed to facilitate the communication of systems that are deployed on diverse platforms.

through a common protocol. Conformance to uniform standards is clearly a barrier in cross-disciplinary contexts since each community tends to develop its own conventions for storing, describing and accessing data.

The BCube Brokering Framework

The BCube project advanced a Brokering Framework by addressing the social, technical and organizational aspects of cyberinfrastructure development. It sought to identify best practices in both technical and cultural contexts by means of engaging scientist with the evolving cyberinfrastructure to achieve effective cross-disciplinary collaborations. The engagement included a number of different communities in guiding and testing the development, with the aim of involving geoscientists at a deep level in the entire process.

BCube adapted a brokering framework that had been developed for the EuroGEOSS project (Vaccari, et al., 2012) and subsequently deployed in the Global Earth Observation System of Systems (GEOSS). Called the Discovery and Access Broker, or DAB (Nativi, et al., 2013), it has successfully brokered millions of data records from dozens of data sources. Guided by the recommendations laid out in the Brokering Roadmap (Khalsa, et al., 2012), BCube sought to demonstrate how brokering could enhance cross-disciplinary data discovery and access by having scientists from different fields create real-world science scenarios that required the use of data from diverse sources.

The approach that BCube promoted was one in which the broker was taught to interact with each community's conventions, allowing the participating systems to interact without adopting a common set of standards. BCube developers then set about configuring a cloud-based version of the DAB to access these sources. This required developing software components, called "accessors", that interacted with each data source. At the start of the project we believed that the suite of accessors that had already been developed for GEOSS could in many cases be reused for brokering the datasets identified in BCube's science scenarios.

The brokering framework is depicted in **Figure 1**.

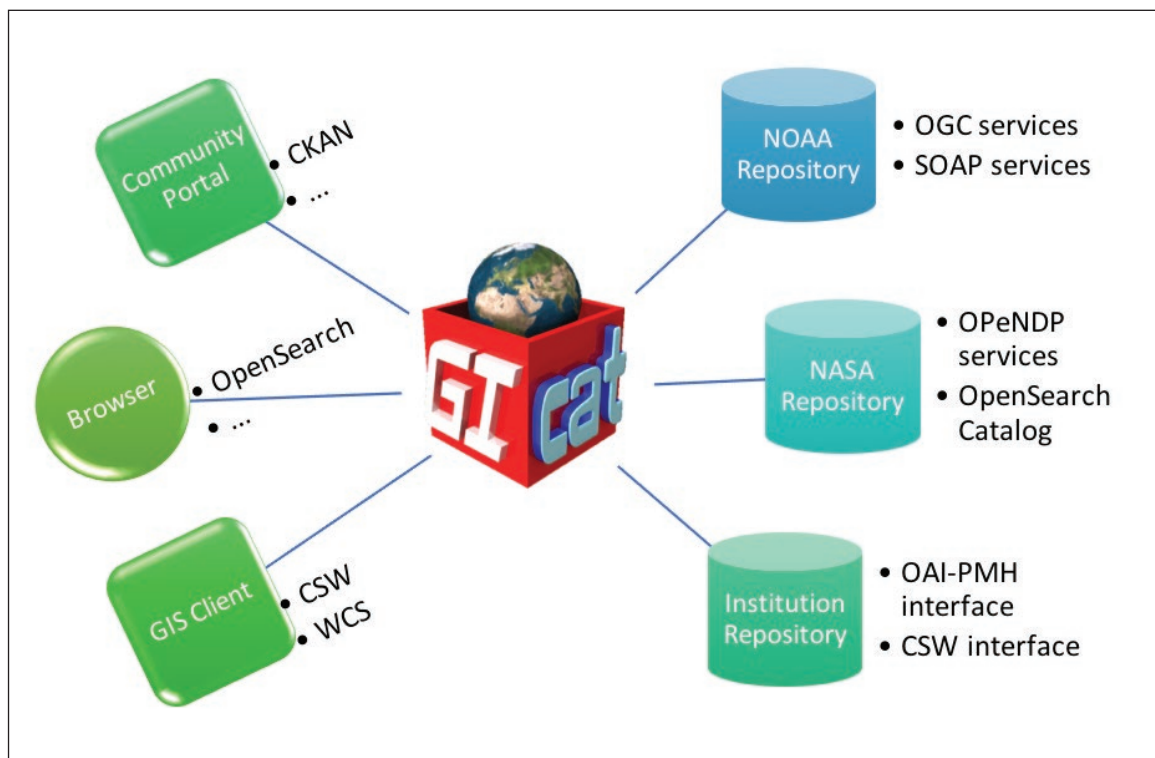


Figure 1: The BCube Broker, based on GI-cat and related software from CNR², mediates two-way requests and responses between clients, depicted on the left and data repositories, depicted on the right, for data query, access, and transform services.

² CNR, Consiglio Nazionale delle Ricerche (National Research Council of Italy), Institute of Atmospheric Pollution Research. <http://www.iaa.cnr.it/>.

Science Scenarios

The project was guided by science scenarios developed by the geoscientists on the BCube team. These scenarios were used to define requirements for the Broker development while engaging the geoscience community with EarthCube. They also provided the basis for evaluating the added value of brokering.

The term “Science Scenario” was used in place of what is more commonly known in software development as “Use Case”. This was in response to a concern in that EarthCube should be solving real, rather than hypothetical problems.

The science scenarios, coming from the fields of hydrology, oceanography, polar science and climate/weather, focused on the specific research needs of each scientist. For each scenario, a team composed of domain scientists and computer scientists was convened to investigate the ability of the BCube Brokering Framework to meet the identified needs of the scientists. These needs determined what new or modified mediation functions the Broker needed to perform in order to fulfill the scenario.

Several different types of scenarios were defined. There were scenarios that described high level science research or education goals without referencing specific data and services. The enactment of these scenarios involved both discovery and access as part of the scenario. The primary type of BCube scenario was the detailed science or education scenario in which the scientist identified specific data sources and services that they wished to have access to. Each scenario described the end-to-end activities required to achieve a science objective. By observing how the objective was accomplished first without brokering and then with brokering we were able to evaluate how the broker was saving time and effort. The flow for this type of scenario is depicted in **Figure 2**.

The third type of scenario the project defined involved configuring the broker to access the resources of a major data repository, thereby making its resources discoverable and accessible, thereby supporting cross-discipline research.

The BCube Brokering Framework gives access to 17 different data repositories serving over 5 million datasets, as show in **Table 2**.

Metadata Brokering

People and autonomous agents find resources, by which we mean data, models, computational services and the like, through the encoded information describing those resources, i.e. metadata. Metadata should also describe how resources are structured and accessed. In brokering a resource, the broker must first access and translate the available metadata and map it to a common internal data model. To serve a metadata record in response to a request the broker maps from the internal model to the model conforming to the protocol consistent with the request.

Unlike the unstructured metadata that supports free text queries used with general search engines, metadata that has been mapped to a common data model enables search by specific features of the data such as its temporal coverage and spatial extent.

The BCube Brokering Framework was already equipped to understand many common protocols and metadata standards, such as OAI-PMH and OGC’s CSW. For some of the services required by the science scenarios, however, a one-time manual mapping was required when the metadata model of resource was not already known to broker. The internal data model of the broker is based on the ISO 19115 family of metadata

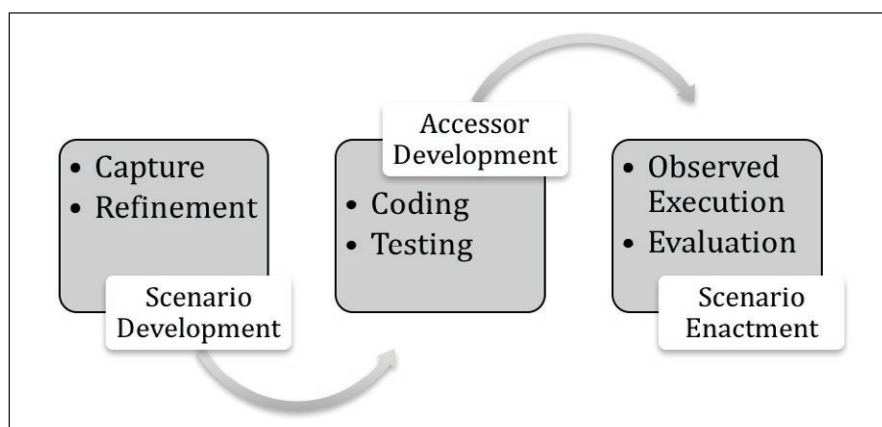


Figure 2: Flow in development and enactment of BCube science scenarios.

Repository/source	Protocol	Number of Datasets
AVHRR SST	THREDDS	62,777
BCO DMO	SPARQL	10,702
Global Multi-Resolution Topography (GMRT)	OGC WMS	13
IRIS Event	Custom	4,213,828
IRIS Station	Custom	544,991
Integrated Marine Observing systems	OGC CSW	601
NASA ASTER	OPeNDAP	22,684
NERRS	SOAP	329
NSIDC	OpenSearch	161
One Geology	OGC CSW	438
PANGAEA	OAI-PMH	356,943
RTOF Models	GrADS	46
Rutgers ERDDAP service	OPeNDAP	1,200
SRTM NASA	OPeNDAP	14,282
UNAVCO GPS	Custom	1,739
UNAVCO SSARA	SOAP	2,000
US NODC	OGC CSW	29,840

Table 2: Resources brokered by the BCube Brokering Framework, along with the access protocol and number of records for each.

standards, but is extensible to accommodate unique community requirements. A basic form of semantic mediation was possible with the BCube broker through augmentation of terms in keyword searches (Santoro et al., 2012).

Data Brokering

If a resource used an access protocol and data format that was already known to the broker it should be simply a matter of pointing the broker to the service endpoint. However, it is common to encounter a service endpoint that does not completely conform to the declared protocols and standards, necessitating customization of the accessor. Customization was always required for protocols or encodings for which there was not an existing accessor. One of the main activities and resource drains within the BCube project was the development and testing of accessors. Midway through the project, an Accessor Development Kit (ADK) was released to help the developers who had been tasked with writing accessors.

Sustainability

The large-scale physical infrastructures (water, power, communication networks) that societies depend on are seen as the responsibility of government and commerce. The internet, which began as a research infrastructure supported by communications protocols, has evolved into a vast, unstructured information resource, enabled through the standards that underlie the World Wide Web. EarthCube, which must build on existing cyberinfrastructure technologies, itself must find a means to become self-sustaining. While it is expected that most of the individual elements developed with EarthCube funding eventually find a means of self-support, it has been the belief within the BCube project that certain foundational elements would need continued funding. In 2001 NSF foresaw the need for such foundational infrastructure and established the NSF Middleware Initiative to define, develop and support an integrated national middleware infrastructure. The focus then was on grid and high-performance computing, and on identify and access management tools (Sun and Blatecky, 2004).

Team members from BCube initiated a working group within the Research Data Alliance, to explore solutions for the governance and sustainability of middleware. In their report (RDA, 2015) “Sustainable Business Models for Brokering Middleware to support Research Interoperability” the Working Group concluded that

the strongest model for sustainment would be one where a federally funded data facility provided guardianship at the stage where the broker was being established followed by a Consortium model and/or Software-as-a-Service model as the broker matured. This approach was anticipated by Ribes and Finholt, (2009) who predicted that in the face of short-term funding, cyberinfrastructure projects will attempt to transition to facilities by forming alliances with the persistent institutions of science in their domain fields.

Lessons Learned

From the start we realized that BCube was not primarily about software development. It was about demonstrating an approach to the construction of EarthCube that would achieve the maximum buy-in from the geosciences community. We aimed to do this by making it easier for geoscientists to find, use and share data and knowledge in an interdisciplinary context without requiring the providers and consumers of that data and knowledge to do extra work. The technical aspects of this were straightforward: write code that mediates the interactions between a distributed and diverse set of clients and servers. The software that was developed, however, had to fit into the greater context of EarthCube. Also, the resources needed to create and maintain this software had to be weighed against other investments necessary for a viable cyberinfrastructure. Furthermore, many researchers felt that the investments being made in technology projects were siphoning off money that should be going to basic research. Until a technology makes it substantially easier to do their work, or opens up opportunities to make new discoveries, scientists will be reluctant to support the EarthCube enterprise.

Factors influencing attitudes towards brokering can also be viewed from the perspective of data providers who wish to fulfil the expectation or obligation of making their data available to users outside their normal clientele. They would support a brokering service only if they had confidence that the service would have long term support and be able to adapt to any changes over time in the provider's data and mission, as well as support the evolving demands of both its customary as well as external, cross-disciplinary users.

These tensions were clearly called out in a report by an EarthCube Advisory Committee (EarthCube, 2016). The Committee felt that EarthCube lacked clear definition and had yet to deliver on its promises. They saw a need for a succinct implementation plan. Funded projects, most of which have yet to move beyond pilot demonstrations, responded that infrastructure development is a long process and that patience is required (Witze, 2016). The need for patience is well articulated in Ribes and Finholt, (2009) who argue that infrastructure development is an occasion for the *long now* – the collapsing of the demands of immediate design and deployment with the work of maintenance and sustainable development.

One of the key lessons that we learned in BCube was that timely delivery of features is vital to keeping scientists engaged during the development phase. Because there were only a small number of software engineers who had access to the BCube source code, and they were doing all the coding and testing of accessors, it was often that a month or longer between the time a scientist who was working with the broker found problems and the time those problems were fixed. Research programs are hard to maintain under such interruptions.

The creation of the Accessor Development Kit (ADK) was meant to mitigate these problems, but the kit itself was new and needed some refinements after initial use, leading to further delays in getting necessary functionality into the broker. Furthermore, developers are hindered in writing robust code to an interface when they have incomplete knowledge of what happens on the other side of the interface, especially when the accessor they are building is invoking operations on data that will be executed by code that they have no access to.

A principal take-home message from the experience of brokering in the BCube project is that in order to achieve the level of community participation in the development and use of software that is intended to become infrastructure, the entire code base should be accessible to developers. Also, since the data sources that are of most interest to geoscientists are often in an evolving, dynamic state, it is challenging to build mediators for them, which argues all the more strongly for open source code that would give the systems developed the responsiveness and flexibility to remain relevant to scientists' needs and interests.

Conclusions

The BCube project successfully demonstrated that it was possible to build a brokering framework that mediated the interactions between clients and servers, where clients could be individuals using a web portal, desktop application software, clearinghouses or other service consumers, and servers were data catalogues, data repositories, and data services. Mediation allowed these clients and servers to each use their own distinct

protocols, semantics, and data syntaxes in managing their data yet still be part of a larger interoperable system, all without needing to install new software or change the way they carried out their operations and workflows. However, the degree of engagement with the science community that BCube sought fell far short of what was hoped for. Delays in delivering functionality were largely responsible for this. In some cases this was compounded by the small amount of time that scientists who signed on to the project had committed. It became clear that an independent interoperability solution based on middleware was viable only if communities become involved in supporting software development and maintenance.

It can be said that new data services can be considered infrastructure only after the users of the technology adapt their behaviours to these new capabilities. EarthCube has yet to deliver the capabilities that would lead to widespread changes in the way geoscientists do their work, but this, indeed, takes time.

Acknowledgements

The BCube team included dozens of people from several different institutions. The author gratefully acknowledges their contributions to achievements of the project. The team composition can be view at: <http://nsidc.org/informatics/bcube/communities>.

Competing Interests

The author has no competing interests to declare.

References


- Bowker, G C, Baker, K, Millerand, F and Ribes, D** 2010 Toward information infrastructure studies: Ways of knowing in a networked environment. In: *International Handbook of Internet Research*. Springer, pp. 97–117.
- Delaney, C, Alessandrini, A and Greidanus, H** 2016 Using message brokering and data mediation on earth science data to enhance global maritime situational awareness. Presented at the IOP Conference Series: *Earth and Environmental Science*, IOP Publishing, p. 12005. DOI: <https://doi.org/10.1088/1755-1315/34/1/012005>
- EarthCube** 2016 EarthCube Advisory Committee Report. Available at: https://www.earthcube.org/sites/default/files/doc-repository/earthcube_rsv_report_final_16_03_21.pdf (accessed 30 October 2016).
- Edwards, J L, Lane, M A and Nielsen, E S** 2000 Interoperability of biodiversity databases: biodiversity information on every desktop. *Science*, 289: 2312–2314. DOI: <https://doi.org/10.1126/science.289.5488.2312>
- Edwards, P N, Jackson, S J, Bowker, G C and Knobel, C P** 2007 Understanding infrastructure: Dynamics, tensions, and design. Arlington, VA: National Science Foundation. Available at: <http://hdl.handle.net/2027.42/49353>.
- Khalsa, S, Parsons, M, Duerr, R, Pearlman, J, Pearlman, F, Browdy, S, Nativi, S, Robinson, E and Dominico, B** 2012 Brokering for EarthCube Communities: A Road Map. National Snow and Ice Data Center. DOI: <https://doi.org/10.7265/N59C6VBC>
- Nativi, S, Craglia, M and Pearlman, J** 2013 Earth Science Infrastructures Interoperability: The Brokering Approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. DOI: <https://doi.org/10.1109/JSTARS.2013.2243113>
- RDA** 2015 Sustainable Business Models for Brokering Middleware to support Research Interoperability: A Report from the Sustainable Business Models Team to the Brokering Governance Working Group of the Research Data Alliance (RDA). Available at: <https://www.rd-alliance.org/group/brokering-ig-brokering-governance-wg/outcomes/sustainable-business-models-brokering-middleware> (accessed 30 October 2016).
- Ribes, D and Finholt, T A** 2009 The long now of technology infrastructure: articulating tensions in development. *Journal of the Association for Information Systems*, 10: 375.
- Santoro, M, Mazzetti, P, Nativi, S, Fugazza, C, Granell, C and Díaz, L** 2012 Methodologies for augmented discovery of geospatial resources. In *Geographic Information Systems: Concepts, Methodologies, Tools, and Applications*, 305.
- Simons, R and Mendelssohn, R** 2012 ERDDAP-A Brokering Data Server for Gridded and Tabular Datasets. Presented at the AGU Fall Meeting Abstracts, p. 1473.
- Star, S L and Ruhleder, K** 1996 Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information systems research*, 7: 111–134. DOI: <https://doi.org/10.1287/isre.7.1.111>

- Sun, X-H** and **Blatecky, A R** 2004 Middleware: the key to next generation computing. *Journal of parallel and distributed computing*, 64: 689–691. DOI: <https://doi.org/10.1016/j.jpdc.2004.03.002>
- Vaccari, L, Craglia, M, Fugazza, C, Nativi, S** and **Santoro, M** 2012 Integrative Research: The EuroGE-OSS Experience. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. DOI: <https://doi.org/10.1109/JSTARS.2012.2190382>
- Witze, A** 2016 Effort to wrangle geoscience data faces uncertain future. *Nature*, 538: 303 (20 October 2016). DOI: <https://doi.org/10.1038/538303a>

How to cite this article: Khalsa, S J S 2017 Data and Metadata Brokering – Theory and Practice from the BCube Project. *Data Science Journal*, 16: 1, pp.1–8, DOI: <https://doi.org/10.5334/dsj-2017-001>

Submitted: 31 October 2016 **Accepted:** 04 January 2017 **Published:** 12 January 2017

Copyright: © 2017 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 