Wagging the long tail of earth science: Why we need an earth science data web, and how to build it

Ian Foster, Daniel S. Katz, Tanu Malik
Computation Institute
University of Chicago

Peter Fox
Tetherless World Constellation
Rensselaer Polytechnic Institute

By adopting, adapting, and applying semantic web and software-as-a-service technologies, we can make the use of geoscience data as easy and convenient as consumption of online media.

Consider Alice, a geoscientist, who wants to investigate the role of sea surface temperatures (SSTs) on anomalous atmospheric circulations and associated precipitation in the tropics. She hypothesizes that nonlinear dynamics can help her model transport processes propagated long distances through the atmosphere or ocean, and asks a graduate student to obtain daily weather, land-cover, and other environmental data products that may be used to validate her hypothesis.

Like the vast majority of NSF-funded researchers (see Table 1), Alice works with limited resources. Indeed, her laboratory comprises just herself, a couple of graduate students, an undergraduate, and a technician. In the absence of suitable expertise and infrastructure, the apparently simple task that she assigns to her graduate student becomes an information discovery and management nightmare. Data are either not available or are of poor quality. Downloading and transforming datasets takes *weeks*. Alice then faces new challenges. Will these new data enrich her compute-intensive model, or simply propagate errors? Or should they seek other, higher-resolution datasets? What software can she use to help answer these questions? We cannot blame Alice if she ultimately abandons this promising avenue of research.

Contrast Alice's experience that evening at home, as she seeks to relax with a movie. She enters a few keywords in a Web browser. The Web integrates distributed sources and discovers deep information to present a wide range of choices; it can even keep her updated of new information, if she subscribes to an alert mechanism. In effect, the Web helps her transform a vast amount of information into knowledge and actionable intelligence, and to pick and choose what is useful from what is not. And once she identifies a suitable movie, it streams reliably to her chosen playback device. She can also, if she so desires, share her experience easily with her friends via email and social networking tools.

Alice's experience as a consumer of online entertainment differs from her experience as an earth scientist for many reasons, but two are particularly noteworthy. First, the consumer industry uses sophisticated "cloud" services to offload the vast majority of the complexity associated with managing, indexing, searching for, and delivering digital content, so that Alice needs only a simple media streaming device in her home. Second, those digital content are described, indexed, and discoverable via a rich collection of mechanisms that may include keywords, reputation, recommendations from friends, and free text search.

Table 1: Heidorn [11]'s analysis of NSF grants over $500 in 2007 shows that 80% of funding was for grants of $1M or less, and that those grants constituted 98% of awards. The majority of NSF-funded research occurs within "long tail" laboratories—as does, we imagine, the majority of both data analysis and data generation.

Total Grants over $500	12,025 ($2,865,388,605)	
80/20 by grant numbers	20% by number of grants	80% by number of grants
Number Grants	2404	9621
Total Dollars	$1,747,95,7451	$1,117,431,154
Range	$38,131,952 - $300,000	$300,000 - $579
80/20 by grant $$	20% by total value = $573,077,721	80% by total value = $2,292,310,884
Number of grants	254	11,771
Range	$38,131,952 - 1,034,150	1,029,9984 - $579

We argue that if Earth Cube is to succeed in its ambitious goal of transforming how earth sciences research is performed, it must deploy comparable technologies to address the needs of "long tail scientists" such as Alice. That is, it must leverage both cloud and semantic Web technologies to make the experience of discovering, using, and integrating geoscience data as convenient as consuming online media. To this end, we must undertake a deliberate effort to assemble **a distributed, discoverable, dynamic web of scientific data for Earth scientists** and to create an **outsourced software-as-a-service (SaaS) infrastructure** to provide the services required for this dynamic earth science web to function and prosper.

The Exaflood and Need for Knowledge Management Infrastructure

Geoscience faces an 'exaflood' of data; in climate science, hundreds of exabytes are expected by 2020 (*Challenges in Climate Change Science and the Role of Computing at Extreme Scale*) [4]. Such data will only be usable by long tail researchers such as Alice if we can provide them with means of overcoming the barriers of data complexity, method heterogeneity, and the inherently multi-scale nature of geoscience.

The last decade witnessed some vital data barriers being crossed. "Big iron" data infrastructures such as Earth System Grid [26][27] and EOSDIS [15] have benefited geoscientists tremendously. Their stable software pipelines make large volumes of homogenous data available to many, and strong governing bodies ensure repeatability and auditability. Yet despite these successes, geoscientists such as Alice still face nightmares because information-rich answers to their data questions remain unavailable. Consider questions, such as (1) "How can 'point' data be reconciled with various satellite—e.g., swath or gridded—products?"; (2) "How is spatial registration performed?"; (3)"Do these data represent the 'same' thing, at the same vertical (as well as geographic) position or at the same time, and does that matter?" Geoscientists have traditionally found answers to such questions embedded deep in scientific publications. Publications, however, provide information about small amounts of data, often giving rise to even more questions.

To navigate such barriers we require a cross-disciplinary knowledge management infrastructure (*Climate Knowledge Discovery Workshop Report*) [5] that combines features of both "big iron" data management and publication infrastructures. To create such an infrastructure, one must pay close attention to key issues that remain traditionally simplified or underrepresented in either infrastructure. These include:

- Data and knowledge are **distributed** over thousands of **disparate** resources, embedded in community-specific databases and encoded in scientific publications scattered in many journals, articles, and websites. These data sources have independent syntax, access mechanisms, and metadata, resulting in considerable heterogeneity.
- Much geoscientific data is hard to **discover,** i.e., to locate and identify. Scientific data shares the same woes as deep web data, which is estimated to be 600 times larger in size than surface web data. Increasingly generated automatically and processed in several data management layers, scientific data resides in isolated data islands that are hard to search and locate in a coordinated fashion.
- The **data-knowledge latency lag** is getting higher. Scientific data doubles every 12 months, which is often faster than it can be converted into useful knowledge. This lag impacts analysis, especially in critical areas such as tsunami-related oceanography and volcanic eruption-related earth sciences.

To effectively represent such aspects of data in a cross-disciplinary knowledge management infrastructure, **we envision an Earth science data web that is as ubiquitous, discoverable, and up-to-date as the commercial Web is to an average person—an Amazon-like (or Yelp-like or Netflix-like) marketplace for science data**. By being interoperable and interconnected, such a scientific Earth data Web will enable rapid discovery and analysis of large quantities of relevant scientific data. And because it outsources hard management, linking, indexing, search—and ultimately analysis and collaboration—tasks, this infrastructure can allow long tail scientists to apply this data without local infrastructure and expertise.

This white paper describes the core technology needed to realize our vision, and addresses issues of scale, governance, and trust necessary for creating the next generation of Earth sciences cyberinfrastructure (CI).

A Scientific Earth Data Web

Scientific data sources are often not constructed with the intention of linking them to a larger entity, such as a scientific data web. Consequently, in constructing such a data web, we must first **pay the cost of integration** as we seek to "understand the data," i.e., work to improve integrity by determining metadata for each input data source, the constraint rules that govern the data instances, and, perhaps most importantly, the rules that describe the relationships between data from different sources.

Internet-scale integration cost mandates cost-effective technology and scalable software tools. It also mandates a scalable governance structure that integrates data from the "long-tail of geoscience" [1]. Our choices affect the geoscientist' experience and their trust in the data Web. We believe that Internet-scale integration requires open standards to democratize the integration process by providing equality to every actor and unrestricted choices in the information science field [23,25].

We argue for (i) Semantic Web as the integrative technology and (ii) Software as a service (SaaS) as the scalable delivery model for an Earth Data Web. The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. The open standards URI [34], RDF [24], and OWL [18] can provide a single, unified view of data across geoscience applications and allow for precise information retrieval.

The SaaS delivery model [26] allows a provider (e.g., Google Docs, Salesforce.com) to run a single version of its software, which many users can access over the network using simple and intuitive Web 2.0 interfaces. Economies of scale mean that the cost per user is much less than if the user tried to provide the function on their own. SaaS has proven its worth in allowing small and medium businesses (SMBs) to slash operational costs, while simultaneously improving their performance through access to higher quality software, by outsourcing payroll, email, accounting, web presence, customer relationship management, etc.

Our technology and delivery choices have been applied effectively in the geosciences:

- Semantic data integration has led to distributed, discoverable platforms such as the **Virtual Solar Terrestrial Observatory** [14,17]. This Observatory integrates data across interdisciplinary collections and is used to evaluate hypotheses, such as exploring the connection between volcano emissions and effects on atmospheric air quality. Semantic approaches have also gained momentum in eScience areas [13], such as solar terrestrial physics, ecology, ocean and marine sciences, healthcare, and life sciences and are now leading to general capabilities such as the Semantic eScience Framework [34].
- **Globus Online** [10,12] is a SaaS research data movement system that allows users to request the movement or synchronization of datasets between remote storage systems that are (most commonly) located in different administrative domains. Web, REST, and command line interfaces allow for both interactive use and integration with application tools and workflows. In just ten months, it has acquired more than 2,000 registered users from physics, geosciences, and biology, among other domains; has been used to move hundreds of millions of files and hundreds of Terabytes of data; and has been adopted by a range of major facilities and science projects as their data movement infrastructure.

To realize a distributed, discoverable Internet-scale scientific Earth data web that is affordable for all, we envision a cloud-based environment for the storage, management and querying of geoscience linked data. Such an environment will enable convenient on-demand access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction [12,19]. We envision the semantic technology-based software and tools on top of the cloud infrastructure that will 1) consume and connect with linked data from respective cross-disciplinary domains in geosciences, 2) query and find data of greatest importance, and 3) develop hypotheses at a scale appropriate to the increasing volume of new data generation and analysis. We next describe specific technologies needed for construction of such a scientific data web.

1) Large-scale linked data on the Web. A general ontology repository for the geoscience community can help them link their data sets. Such a repository will provide the building block for global search,

browsing, and inference using standard web encodings and protocols. Repository systems exist for publishing and sharing ontologies and vocabularies for content indexing, information retrieval, content integration, and other purposes, e.g., Cupboard [6], BioPortal [20], OOR [2], and ONKI [27]. However, most such systems are from biology, although their adoption has promise for an ontology repository. Further, ontology repositories often result in separate islands, with no connections to other repositories. Thus, we propose an ontology repository derived via a combination of both community-driven annotation [31] and extraction of ontologies from the semantic deep web [32,30].

Ontology repositories face the issue of redundant concepts. To that effect we propose a Universal Semantic Object (USO) identifier. A USO shall be associated with data coming from various data sources. The USO will form the basis for incorporating any incoming ontology, concept, dictionary, and reference data. USOs will form a basis to link data sources globally. The system will permit USOs and other discovered metadata to be cached or instantiated for indexing, query, etc., as in data warehouses today. To enable such querying and indexing, we will however, need powerful computers: a CI that we describe next.

2) Large-scale cloud computing infrastructure for querying and finding relevant information in a semantically rich world. Our ability to query and find interesting and relevant information in the semantic world resides on our ability to harness large-scale computers that can materialize inferences through forward-chaining, materializing views, computing transitive closures so that queries run faster. The Billion Triple Challenge [3] currently explores interesting and even provocative topics, ranging from query rewriting to mobile applications, ontology hijacking, and navigation. Most proposals are for investing in bigger and powerful machines to pay this cost. However, parallel and distributed computing has shown great promise in the business domain, and frameworks such as MapReduce [9] can scale up query processing and reasoning in the semantic Web. These frameworks, although most suitable for embarrassingly parallel problems, have now been adapted at a large scale for graph-based problems and thus can be useful for semantic applications also.

Simply buying machines, whether small numbers of large systems or many smaller ones, is not sufficient to address the scalability problem. It is important to carefully analyze and evaluate the shape and characteristics of the data—often using metrics, summary structures and signatures. Investment in such data structures can lead to building a robust CI that can scale to large volumes of data. Use of such optimization structures has already provided magnitude of performance improvements in querying of biological ontology databases wherein benchmarking analyses are used to analyze and improve the performance of the system as much as possible before resorting to investing in new infrastructure [16].

3) Dynamic Updates of Semantic Web. Increasing amounts of data created via an ontology repository will quickly lead to information deluge. We believe in many sciences this is already true and requires development of applications based on selective information dissemination, wherein data is distributed only to interested clients or have expressed interest in the past through queries. A scientific data web will require use of event-based architectures, such as publish-subscribe (Pub/Sub) systems that can efficiently supply user interests with available information [7]. Pub/sub systems have already started taking ground in some geosciences domains. For instance, the Polar Information Commons [22] adopts a data advertising approach called “data casting”. The GEO Portal3 [8] aggregates geoRSS feeds and NASA has supported further development of data casting for both whole collections and individual files or records.

Casting supports subscriptions to the most recent data. We envision a system of “active queries,” in which users may specify temporal data requirements with their subscription. This will, in effect, allow a scientist to register their interest in future versions of a dataset, for example, new data that is added within a temporal range that could be bounded by “present” or a specific date. Including time-scales in the data feeds through requirement specification can allow a scientist to perform hypothesis evaluation at various time granularities. Accordingly, we propose creation of a geoscience middleware that can keep a scientist’s data of interest synchronized with the original data source via the ontology repository.

Conclusion

Our vision of a scientific data web is already being realized as small-scale efforts in the semantic web community and some individual disciplines. We believe that this vision is compelling, because of 1) the maturity of the underlying research in computer and information systems, 2) the success of scientific data management in other information-rich sciences such as astronomy and biology, and 3) our personal experience in building large-scale scientific systems. We also believe that software-as-a-service (SaaS) provides the vehicle by which we can scale out the deployment and application of this technology to the long tail of earth science; again, we base this statement on experiences both in other fields and from our personal experience. However, for these methods to be fully adopted by the Earthcube community, they need to be adopted by all stakeholders and advanced through community standards and community knowledge.

References

[1] C. Anderson, C. The Long Tail. *Wired Magazine*, 12(10):Retrieved from http://www.wired.com/wired/archive/12.10/tail_pr.html, 2004.

[2] K. Baclawski and T. Schneider, The open ontology repository initiative: Requirements and research challenges, *Workshop on Collaborative Construction, Management and Linking of Structured Knowledge*, 2009.

[3] The Billion Triple Challenge, http://www.cs.vu.nl/~pmika/swc/btc.html.

[4] Challenges in Climate Change Science and the Role of Computing at Extreme Scale. http://science.energy.gov/~/media/ber/pdf/Climate_report.pdf

[5] Climate Knowledge Discovery Workshop Report, https://verc.enes.org/collaboration/attachments/198/Ludwig.pdf

[6] M. d 'Aquin and H. Lewen, Cupboard--a place to expose your ontologies to applications and the community, *The Semantic Web: Research and Applications*, 2009.

[7] P.T. Eugster, P.A. Felber, R. Guerraoui, and A.M. Kermarrec, The many faces of Publish/Subscribe, *ACM Computing Surveys*, 2003.

[8] The GeoPortal, http://www.geoportal.org.

[9] S. Ghemawat and J. Dean, MapReduce: Simplified data processing on large clusters, *Operating System Design and Implementation*, 2004.

[10] Globus Online. [Accessed July 11, 2011]; Available from: www.nersc.gov/users/data-and-networking/transferring-data/globus-online/.

[11] P.B. Heidorn, Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, 57(2):280-299, 2008.

[12] I. Foster, Globus Online: Accelerating and democratizing science through cloud-based services. *IEEE Internet Computing*(May/June):70-73, 2011.

[13] P. Fox and J Hendler, Semantic eScience: Encoding Meaning in Next-Generation Digitally Enhanced Science, *The Fourth Paradigm*, 2.004

[14] P. Fox, D. McGuinness, L. Cinquini, P. West, J. Garcia, and J. Benedict, Ontology-supported scientific data frameworks: The virtual solar-terrestrial observatory experience, *Computing in Geoscience*, Vol. 35, 2009.

[15] B. Kobler, J. Berbert, P. Caulk and P.C. Hariharan, Architecture and design of storage and data management for the NASA Earth Observing System Data and Information System (EOSDIS), *Mass Storage Systems*, 1995.

[16] P. LePendu, N. F. Noy, C. Jonquet, P. R. Alexander, N. H. Shah, and M. A. Musen. Optimize First, Buy Later: Analyzing Metrics to Ramp-up Very Large Knowledge Bases, *The International Semantic Web Conference*, 2010.

[17] D. McGuinness, P. Fox, L. Cinquini, P. West, J. Garcia, J. L. Benedict, and D. Middleton, The Virtual Solar-Terrestrial Observatory: A Deployed Semantic Web Application Case Study for Scientific

Research, *AI Magazine*, Vol. 29, 2007.
[18] D.L. McGuinness, F. Van Harmelen and others, OWL web ontology language overview, W3C recommendation, 2004.
[19] P. Mell and T. Grance, The NIST definition of cloud computing, National Institute of Standards and Technology, 2009.
[20] N. Noy, N. Shah and et. al, BioPortal: ontologies and integrated data resources at the click of a mouse, *Nucleic acids research*, 2009.
[21] M. Parsons, O. Godoy, E. LeDrew, T. F. deBruin, B. Danis, S. Tomlinson and D. Carlson, A Conceptual Framework for Managing Very Diverse Data for Complex, Interdisciplinary Science, *Journal of Information Science*, 2011.
[22] The Polar Commons, http://polarcommons.org/.
[23] Principles for open data in science, http://pantonprinciples.org/
[24] Resource description framework (RDF) model and syntax, World Wide Web Consortium, http://www. w3. org/TR/WD-rdf-syntax.
[25] TEDx Open Science talk by Michael Neilson, http://michaelnielsen.org/blog/open-science-2/
[26] M. Turner, D. Budgen and P. Brereton. Turning Software into a Service. *IEEE Computer*, 36(10):38-44, 2003.
[27] O. Valkeapaa and O. Alm and E. Hyvonen, Efficient content creation on the semantic web using metadata schemas with domain ontology services, *The Semantic Web: Research and Applications*, 2007.
[28] D.N. Williams, R. Ananthakrishnan, et al., The Earth System Grid: Enabling Access to Multi-Model Climate Simulation Data. In: *Bulletin of the American Meteorological Society*, 90(2):195-205, 2009.
[29] D.N. Williams, D.E. Bernholdt, I.T. Foster and D.E. Middleton, The Earth System Grid Center for Enabling Technologies: Enabling Community Access to Petascale Climate Datasets. In: *Cyberinfrastructure Technology Watch (CTWatch) Quarterly*, 3(4), 2007.
[30] V.R. Benjamins, J. Contreras, O. Corcho and A. Gomez-Perez, Six challenges for the semantic web, *Knowledge and Reasoning*, 2002.
[31] J. Geller, S.A. Chun and Y.J. An, Toward the Semantic Deep Web, *Computer*, Vol: 41, 2008.
[32] S. Handschuh and S. Staab, Annotation for the semantic web, *IOS Press*, 2003.
[33] URI, http://www.w3.org/Addressing/#background
[34] Semantic eScience Framework, http://tw.rpi.edu/web/project/SeSF