ADDRESSING DATA ACCESS NEEDS OF THE LONG-TAIL DISTRIBUTION OF GEOSCIENTISTS

Tanu Malik, Ian Foster

Computation Institute
University of Chicago
tanum@ci.uchicago.edu, foster@anl.gov

ABSTRACT

Data and computation are fundamental to advances in geoscience research and discovery. However, geoscientists currently spend too much time looking for the "right" data, subsequently accessing these data, and then transforming them into a form suitable for analysis. This data management overhead affects a scientists' competitive advantage in making useful contributions. Several cyber-infrastructure (CI) efforts are being undertaken to improve the data management needs of the long-tail geoscientists. In this paper, we highlight characteristics of CI solutions that will form the basis for the successful and widely adopted solutions.

Index Terms— Cyber-infrastructure, Long-tail of science, Data Management, Globus, publishing tools

1. INTRODUCTION

Geoscientists must increasingly consider data from multiple disciplines and make intelligent connections between the data in order to advance research frontiers in mission critical problems, such as assessing the rise of sea surface temperatures, modeling geodynamical earth systems from deep time to present, and examining in detail the causes and consequences of global climate change [12]. As a first step towards making timely and relevant connections, scientists require data and resource access, made available through simple and efficient protocols and web services, which allows them to conveniently transmit, acquire, process and inspect data and metadata.

The last decade witnessed some vital data and resource access barriers being crossed. "Big iron" data infrastructures, such as Earth System Grid [21] and EOSDIS [14] enabled geoscientists with large volumes of simulation and observational datasets; Protocols, such as the OPeNDAP suite [16] make data access convenient; and strong governing bodies, such as those developed by the Open GeoSpatial Consortium (OGC) [17] ensure standards for interoperability, repeatability and auditability. All this remarkable growth in access, however, addresses needs of publishers of large data and ignores consumers of that data. To-date limited access mechanisms exist for the consumers, who fetch subsets, analyze them, and, more often than not, generate new data and analysis, which finally gets published in scientific articles.

While several cyber-infrastructure approaches to have been proposed recently [4,6,13] to address the needs of the consumers, in this paper, we highlight that the success and adoption of such approaches hinges on their ability to include and serve the needs of those consumers who lie on the long-tail distribution of geoscience research. To make the case for those on the long-tail distribution, we first define geoscientists who fall in this category. We then describe data management challenges that such geoscientists face in conducting their research. These challenges emerge from finding appropriate datasets, have access to resources and tools to manage and understand the data sets, and finally to derive knowledge from the datasets. We then present qualifying characteristics of solutions that will continue to serve the needs of these scientists in the long-term. Finally, we present our own efforts in building such solutions and give a vision of the future CI in which such solutions can be useful.

2. THE LONG-TAIL DISTRIBUTION OF SCIENTISTS

The long-tail distribution is characteristic of problem domains in which a larger share of population rests with the tail of the probability distribution [1]. We attribute this long-tail distribution to the vast majority of NSF-funded researchers (see Table 1) who work as single investigators and manage a small or medium sized laboratory consisting of a couple of graduate students, an undergraduate, and maybe a technician.

978-1-4673-1159-5/12/$31.00 ©2012 IEEE

Table 1: Heidorn [5]'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

While Table 1 shows an aggregated trend, the long tail behavior is evident in individual sciences, including Geoscience.

The long tail distribution of single investigator research laboratories suffers because of the inherent mismatch in research expectation and the available expertise to carry out all the data and information tasks related to the research. Consider the mission-critical problem of investigating the role of sea surface temperatures (SSTs) on anomalous atmospheric circulations and associated precipitation in the tropics [2]. Such a task inevitably requires making data connections between diverse disciplines of geoscience. To make these connections effectively, a geoscientist must:

- **Discover data,** i.e., locate and identify, geoscience data, such as daily weather, land-cover, and other environmental data products, from several distributed, disparate data resources. Often these data are embedded in community specific databases, but often are not readily available since they are 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.

- **Conduct data management tasks,** such as downloading, storing, transforming large datasets into compatible formats and from several distributed sources. While this requires investment of time, it also requires adequate resource provisioning and maintenance of the resources on the consumer-side over time.

- **Make data interoperable.** Most scientific data are not constructed with the intention of linking them to other datasets. In our example case, the geoscientist has to make connections, such as (a) Reconcile 'point' data with various satellite, e.g., swath or gridded products; (b) Determine how spatial registration is performed, and (c) Determine if data represent the 'same' thing, at the same vertical (as well as geographic) position or at the same time. To answer these information-rich questions, data must first be made interoperable, a task which is currently achieved manually by determining metadata for each input data source, finding the constraint rules that govern the data instances, and, perhaps most importantly, determining the rules that describe the relationships between data from different sources [11]. It is not surprising to find that several investigators, at the expense of research time, pay this cost of understanding the data.

- **Reliably extract features of interest** and knowledge from the massive archive of multi-dimensional datasets, many of which exhibit a large degree of spatiotemporal sparseness. This requires readily available computational tools and data mining algorithms to extract the fullest information from the datasets.

Given the lack of suitable infrastructure and often expertise, singly investigator labs become more prone to taking ad-hoc or inefficient approaches in conducting these tasks. In the long run, this approach towards data management fundamentally challenges the competitiveness of a long tailed geoscientist [9].

3. CHARACTERISTICS OF A DATA ACCESS PLATFORM

Several approaches [4,6] can be used to address the data access needs of the long tail of the geoscientists. In this paper, our objective is to not provide a complete data access infrastructure solution or rate solutions, characteristics that any solution for geoscience must possess in order to truly and effectively address the needs. This feature list emanates, partly, from our experience in addressing data access needs of scientists from several domains, and partly, from discussions held with geoscientists as part of the EarthCube [3] Data Access Workshop [7], which was coordinated by the authors. Our second objective is to list exemplar solutions that possess several of these characteristics, and are effectively addressing some of the needs.

To effectively address the needs of the long tail distribution of geoscientists, we need cyber-infrastructure that is:

A. **Based on SaaS paradigm**: The SaaS delivery model [20] 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. The model is transformative for the long tail of scientists since it obviates the need for installing, operating and maintain sophisticated cyber-infrastructure. The model results in economies of scale, which means that the cost per user is much less than if the user tried to provide the function on their own.

B. **Architected as a platform**: The long tail, by virtue of its length consists of a wide variety of users. While, most of these users need consistent systems, which have well defined entry and exit points for storing, replicating and transforming datasets (80% roughly), the remaining users need *frameworks*, which have several entries and use points to addresses personal information needs. [12]. To democratize the process of data management including the long tail, we propose use of platforms to provide a variety of tools for data management and thus provide combined benefits of both systems and frameworks.

C. **Rated by user experience**: In our experience with GO, we have begun to consider usability studies and user experience as a dominant force behind CI adoption. Simple and intuitive Web 2.0 interfaces of GO have engaged users and incentivized them to use the system. We consider user-friendly designs as a fundamental component that makes discovering, using, and integrating geoscience data convenient and pleasurable.

D. **Turns raw data into smart content**: Manual procedures for understanding data are redundant and form the most time-consuming tasks. We need semantic technologies that use well-defined vocabularies and ontologies to pull in undefined data and push out defined data with the proper meaning attached in the form of new semantically relevant metadata describing the unstructured content.

E. Encapsulates end-to-end theory: In sciences, data is often generated at supercomputers and experimental locations and is shipped for analysis to multiple, remote distributed locations. End-to-end theory eliminates as many middle layers or steps as possible to optimize performance and efficiency in any process.

Based on these characteristics, in this paper we propose, using scalable cloud-computing infrastructures, Web 2.0 interfaces, integrative semantic web technologies, which can contribute towards building a *knowledge platform* that can discover and access distributed, cross-disciplinary data The platform, through the use of open standards and the Software-as-a-Service (SaaS) delivery model will provide a uniform way for geoscientists to perform the majority of their data and information management tasks. The platform will be expansive in that it will allow integration of applications to provide personalized query and search services [10].

4. EXEMPLAR SOLUTIONS

GLOBUS: An important data management need of scientists is moving data between source and destination endpoints. In the past, data movement, has been painfully tedious and time-consuming, requiring configuration issues, firewall issues and unexpected failures. In November, 2010, the Computation Institute launched Globus Online (GO) [8], and within a year the service gathered more than 2500 registered users, moved ~800 TB of user data approximating about 100 million user files, and registered more than 150 national and international endpoints. Users have quoted rave reviews: "*I expected to spend four weeks writing code to manage my data transfers; with Globus Online, I was up and running in five minutes" [7]*. The service has been designed with (A), (B) and (E) as its guiding principles.

Scientific Object and Linking Environment (SOLE): `SOLE` [18,19] is a bibliography tool for linking research papers with products of a scientific experiment, such as source codes, datasets, annotations, workflows, packages, and virtual images. SOLE shares most of the characteristics of CI described earlier. The system is end-to-end since it enables an author to effectively share scientific objects with the final output of research paper for the purpose of reproducible research. It is designed with the user in mind and eases author burden by allowing authors to easily associate URLs with scientific objects through a tagging mechanism and obtain a Web representation for referencing scientific objects in the paper. It turns raw data into smart content, by enriching the research paper with the resulting links. Multiple authoring environments, such as Wiki, Latex, Word, PowerPoint and can include the SOLE metadata file and reference objects in the paper. Readers can explore the scientific objects directly by clicking links in the paper. SOLE is currently a work in progress and given the varying requirements of several domains, we are working towards a platform-based solution.

GeoBase: GeoBase [15] uses end-to-end principles and SaaS paradigm to deliver up-to-data analysis to the users. Most geoscience data resides in self-describing format files, and are supported by slow interfaces for multi-attribute analysis. In GeoBase, we retain the logical view of the data over the spatio-temporal files but use a key-value storage system to enhance performance. Thus GeoBase layers a multi-dimensional index over the key-value store to provide multi-attribute analysis capability by linearizing multi-dimensional location and time information in spatio-temporal datasets into one-dimensional space of the key-value store. This allows for multi-attribute analysis over scalable key-value stores, thus delivering high performance. The key-value stores retain the scientific data model by internally mapping the model to contiguous, low-level byte extents at the file level. This correspondence is crucial to provide an end-to-end system in which file data is not translated into a key-value systems' native format.

5. CONCLUSION

Science is increasingly data driven, computational, and collaborative. Information management needs of the long tail distribution of geoscientists increasingly require sophisticated knowledge management cyber-infrastructure. We have argued that the needs of the long-tail of geoscientist can be effectively addressed by investing in cost-effective, scalable, and integrative technology. Based on our experience in addressing such needs, we have provided real examples. Finally, we have described the architecture of a knowledge platform.

11. REFERENCES

[1] Anderson. C, "The Long Tail" *Wired*, October 2004

[2] Chang, P., Saravanan, R., Ji L. and Hegerl, G., The effect of local sea surface temperatures on atmospheric circulation over the tropical Atlantic sector, *Journal of Climate*, Vol: 13(13), 2000.

[3] Earthcube, http://earthcube.ning.com/

[4] Earthcube Capabilities, http://earthcube.ning.com/page/capabilities

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

[6] EarthCube WhitePapers, http://earthcube.ning.com/page/whitepapers

[7] EarthCube Discovery, Mining and Access group http://sites.google.com/site/earthcubeddma

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

[9] Foster, I., High-performance, collaborative computing for the long tail of GIS, Keynote Talk at *Workshop on High Performance Distributed GIS*, in conjunction with HPDC, 2011

[10] Foster, I., Katz, D., Malik, T. and Fox, P., Wagging the long tail of earth science: Why we need an earth science data web, and how to build it, http://semanticommunity.info/@api/deki/files/13867/=079_Foster.pdf

[11] Fox, P. and Hendler, J., Semantic eScience: Encoding Meaning in Next-Generation Digitally Enhanced Science, *The Fourth Paradigm*, 2004

[12] Fox, P., (Semantic) Data Frameworks—Experiences with Virtual Observatories in Geosciences.

[13] Killeen, T.,Transforming the Conduct of Research and Education in the Geosciences through EarthCube. http://wiki.esipfed.org/images/6/62/Transforming_Geosciences_through_EarthCube.pdf

[14] Kobler, B. and *et. al*, Architecture and design of storage and data management for the NASA Earth Observing System Data and Information System (EOSDIS). In: *Mass Storage Systems*, 1995.

[15] Malik, T., Best, N., Elliott, J., Madduri, R. and Foster I., Improving the Efficiency of Subset Queries on Raster Images., *Second International Workshop on High Performance and Distributed Geographic Information Systems*, 2011.

[16] OPeNDAP, http://www.opendap.org

[17] OGC, http://www.opengeospatial.org

[18] Pham, Q., Montella, R., Malik, T. and Foster I., SOLE: Linking Research Papers with Scientific Objects, *International Workshop on Provenance and Annotation* (IPAW), 2012.

[19] Scientific Object Linking and Embedding SOLE, http://ci.uchicago.edu/sole

[20] Turner, M., Budgen D., and Brereton, P., Turning Software into a Service. IEEE Computer, 36(10), 2003.

[21] Williams, D and *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.