Int J Digit Libr (2007) 7:113–116
DOI 10.1007/s00799-007-0029-2

COMMENTARY

Sharing and caring of eScience data

Chaitanya Baru

Published online: 2 September 2007

Abstract This commentary reflects on the issues presented in this volume from the perspective of a large eScience project, GEON, whose aim is to promote data integration in the geosciences within the US and abroad. Technical, social, and regulatory challenges accompanying the collection, curation, and sharing of eScience data are discussed. Opportunities and barriers to engaging in international eScience collaborations are highlighted.

Keywords Cyberinfrastructure · eScience

Technology has been moving rapidly in the direction of enabling the generation, collection, and storage of vast amounts of data. However, the organization of this data in the form of well-curated digital libraries to enable discovery and reuse can be a very labor and expert-intensive activity. Archiving and preserving the information for long-term use is an even more challenging task, requiring an examination of business models and an evaluation of the "future value" of "current data." The collection of papers in this issue speaks to the need to study and understand user requirements, expectations, and the cultures of use and sharing. In addition, there is the requirement for the right set of technologies—the *cyberinfrastructure*—to implement the necessary tools.

Cyberinfrastructure has been described as "*the comprehensive infrastructure needed to capitalize on dramatic advances in information technology*" and has been recognized as being essential to support the frontiers of research and education.[1] "*E-Science*" refers then to the science enterprise enabled by the use of such cyberinfrastructure and has been defined as, "*Science increasingly performed through distributed global collaborations enabled by the Internet, using very large data collections, terascale computing resources and high performance visualizations*."[2] This definition, formulated in 2003, and focusing on wide geographic distribution and high-end computing and storage, may need to be expanded. A number of emerging *bona fide* e-Science disciplines have projects that may not be as widely distributed (e.g., may only be regional or continental rather than global scale), or may not require very high end computing or storage.

The nature of computational science and informatics is influenced by the information technology environment into which it is born. The physical sciences—physics, chemistry, astronomy—set the early direction in computational science, emerging in the era of mainframe computers, "flat" file systems, and FORTRAN. Indeed, these applications were themselves strong drivers of the technology, leading to the development of FORTRAN itself and vector-oriented hardware. The field of bioinformatics emerged at a time when the Internet was beginning to mature and relational database technology was a relatively recent invention. The first set of human genome centers, funded in the early 1990s, were required to have an "*Informatics Core*." When the center at the University of Michigan acquired Sun workstations and an Oracle database for its Informatics Core, it

C. Baru (✉)
San Diego Supercomputer Center,
University of California, San Diego, USA
e-mail: baru@sdsc.edu

C. Baru
MC 0505, 9500 Gilman Drive,
La Jolla, CA 92093-0505, USA

[1] NSF cyberinfrastructure vision for 21st Century Discovery, http://www.nsf.gov/pubs/2007/nsf0728/index.jsp.

[2] Oxford e-Science center, http://e-science.ox.ac.uk/public/general/definitions.xml.

was considered *avant-garde* and cutting edge. The e-Science areas that are emerging now—such as, geoinformatics, ecoinformatics, archaeoinformatics—arrive at a time of *mashups*, Google Search, mySpace, and YouTube! At the same time, we have developed a sophisticated understanding of information technologies, and live in a so-called "knowledge economy", where the Internet. Digital libraries have also emerged over the past 10 years, since the initiation of the first set of NSF-funded digital library projects in the US in the mid-90s. The field has kept moving forward, with digital content inexorably wiping out printed materials and challenging the very existence of brick and mortar libraries. Indeed, it is interesting to note that the founders of Google were supported as graduate students at Stanford University by one of the original NSF projects on digital libraries.

Over the past couple of decades, computational science and e-Science disciplines have also benefited from decreasing costs, increasing capacity and increasing reliability of hardware—cheaper, faster processors, storage and instrumentation. There has been a rapid increase in our ability to amass data—whether by being able to run larger computer simulations that produce larger outputs, or by deploying sensors of all varieties—from microscopes to telescopes, sensor networks to particle detectors, and satellite platforms to underwater instruments. This has led to a data deluge that is following an exponential growth curve. With e-Science, the expectation is that, in addition to published results, all or much of this "raw" data (and "low level" derived products) would be made easily accessible and easily usable by subsequent users of the information—including subsequent generations of scientists.

The Geosciences Network (GEON, http://www.geongrid.org) provides a useful case study of some issues related to data discovery, access, integration, and provenance. A collaboration involving 16 institutions, GEON is an Information Technology Research (ITR) project focused on "... *the pressing need in the geosciences to interlink and share multi-disciplinary data sets to understand the complex dynamics of Earth systems."*[3] Heterogeneity across sub-disciplines and, therefore, in data sharing cultures, was "built in" to the project, from its very inception. The project PIs represented a range of Earth Science disciplines from geophysics to geology; a range of data types from field data collections, to sensor datasets, experimental data, and simulation outputs; and a range of organizational experience and backgrounds from highly individual PIs, to members of consortiums, and managers of consortiums. The focus on *data integration* introduced a new dimension to the Earth Sciences community, different from field data collection (which is the work of PIs) or hosting of curated data repositories and archives (done by groups and/or institutions that provide a service to the Earth Science community).

The goal of GEON is to be able to serve as a portal for *finding* and *integrating* Earth Science data. There was the usual question of why Google was not sufficient for finding data. Users quickly understood the need for some amount of quality control on datasets and the need for some level of standards for metadata, including for spatial information. They also understood the heterogeneity of terminology that exists across databases and data sets, and the need for "mapping" different databases to standard vocabularies and ontologies. The level of concern was different among different groups. Defining reference ontologies and linking datasets via ontologies, an area where GEON has developed several innovations, was high on the list for geologists compared with geophysicists. There is some familiarity with common data standards in the geophysics community due to a history of data sharing via common data archives such as IRIS. There is more heterogeneity in many geologic data collections. However, there is now increasing emphasis on the need to define standard schemas and ontologies, because of the increasing awareness of the need for data integration. Earth science data are valuable—they are expensive to collect and data sources may be lost (due to urban development, changes in permitting regulations, etc.)—so there is interest in ensuring that existing data are used to the fullest extent possible.

Different groups also have different needs for data integration. One fundamental distinction is in the dimensionality of data—2D (surficial) data versus 3D (underground) data. GEON initially focused mostly on 2D data, though the emphasis has now shifted to full integration of 3D, 4D and, indeed, multidimensional data. The concept of an *OpenEarth Framework* has been put forward for this purpose.[4] With increasing numbers of data repositories and digital libraries coming into existence, there is also now the appreciation of the need for *metadata catalog interoperability*. Users would like search functionality, such as *GEONsearch* in the GEON Portal (http://portal.geongrid.org), to not only search resources registered with the GEON metadata catalog, but also other metadata catalogs, e.g., those at, say, USGS, in order to perform a comprehensive search for data. This approach is already being implemented in a collaborative project between GEON and EarthScope to develop an EarthScope Portal that will provide access to a single *virtual* catalog of all EarthScope data products. A metadata search of the EarthScope portal would issue a search to the metadata catalogs at IRIS, UNAVCO, and Stanford/ICDP—which are the three repositories for primary EarthScope data. A similar collaboration is being discussed between GEON and some groups in the USGS so that a single search in GEON can

[3] GEON: A research project to create cyberinfrastructure for the geosciences,, http://www.geongrid.org/about/GEON1.

[4] GEON 2.0: A cyberinfrastructure facility for data integration in the earth sciences, http://www.geongrid.org/about/GEON2.

reach into catalogs at the USGS as well. For such data portals to work there is clearly the need for establishing interoperability protocols that allow metadata and data search and sharing. In addition, since these portals provide users with a workspace where they can analyze data, run models, and generate new data products, there is also the need to curate such products. This requires the ability to link back to the source data sets that were used in the analysis and leads to a variety of issues in data provenance, persistence of digital objects, and reproducibility. While there are significant technical and organizational challenges in ensuring provenance and reproducibility of results, especially in distributed environments, this remains a high priority requirement for users.

The promise of cyberinfrastructure is that it will ease the burden of accessing (remote) online resources—data repositories, digital libraries and archives, computing power, and various software tools and applications. The theme of sharing and collaboration is implicit—via Grids, virtual organizations, and collaboratories. While technologies can make such sharing easier, the attitude towards sharing, and the nature and level of sharing itself is certainly influenced by the nature of the science and the nature of the problem being studied. Sharing data from a particle detector in high-energy physics may be easier, as indicated by large international collaborations in this area (CERN, the Large Hadron Collider, the Laser Interferometer Gravitational Wave Observatory). It does not matter whether a particle accelerator is located in Switzerland, France, or, for that matter, Texas. The physics is still the same! Similar arguments would apply for telescopes observing the stars and space. A high degree of sharing is also possible, for example, in biomedical applications, where the subject being studied is the same, e.g., humans or rats. In the Biomedical Informatics Research Network project (BIRN, http://www.nbirn.net), for example, there are many examples of success in data sharing of the original data (brain images) among groups working on similar data (e.g., structural MRIs) or the same problem (e.g., Alzheimer's). The Incorporated Research Institutions for Seismology (IRIS, http://www.iris.edu) is a good example of a successful data archive facility. The sharing is centered on a common data type (seismic). In the article by Collins et al. they have observed differences in behaviors and the success of collaborative e-Science environments for different projects and groups of scientists. The importance of iterating with the user community on the design of such systems is highlighted in Tsoi et al.

Sciences that are "location-based," i.e., focused on collecting data and/or modeling a local area—as opposed to regional, continental, or even global-scale—have more difficulty sharing their data. Borgman et al.'s article touches upon this issue. "Little science" projects are by definition confined to local or regional scale. The data may not be collected with larger scale reuse in mind. The real issue appears to be whether there is a global model into which the local data could be placed. For the geologic data in GEON, for example, locally collected geologic data taken from widely dispersed sites across the planet may actually fit together under, for example, the common plate tectonic model. However, comparisons and integration can be made difficult due to lack of standard protocols and terminology. It appears that the protocols and culture of sharing data is even more interesting in fields like archaeology. While the exact location of sites and "digs" may not be revealed for good reason, the actual "raw" data collected may also not be easily shared. This is a field in which the field observations are all important—and form almost the entire basis of any subsequent explanations. Thus, entire models may be built that are heavily dependent on a particular interpretation of the particular data. There are no means for other researchers to "test" the model or recreate the results. There are only alternative interpretations.

As has been amply demonstrated by the open Internet, it is easiest to share opinions—since they generally come with no real data attached! The moment data is to be shared there is need for appropriate metadata. Generating and providing complete metadata is the Holy Grail for facilitating easy discovery, interpretation, and reuse of data (and, for that matter, models). Thus, there is the ever-present need and desire to develop (and adhere to) metadata standards, develop tools for automated generation of metadata, provide annotation tools, and deploy systems for tracking provenance. The articles by Candela et al., Gahegan et al., Hunter et al., and Witt et al. focus on technologies that help users navigate through vast, possibly distributed, digital library collections. Providing good support for such search and navigation requires the tagging of data with the appropriate metadata, from the earliest stages of data generation to the most advanced stages of use. Furthermore, as mentioned before, the real challenge is to ensure reproducibility of results in a distributed environment with multiple administrative domains (i.e., different institutions, departments, etc). So, does all data need to be at a central place, if reproducibility is of paramount interest? Barros et al. describe efficient ways to input data in the field. In addition, scientists also want a number of tools that will work in the field, e.g., to perform basic, low-level quality assurance and quality control (QA/QC), and associate the data with the corresponding metadata at that point. The desire is that such metadata should be comprehensive in describing the original data and information to make it suitable for any subsequent use—or, indeed, prevention of misuse. It should describe not only the "what" (descriptive metadata) but also the "how" (lineage, provenance) and "why" (contextual information). With techniques such as terrestrial laser scanning (TLS) and a variety of field sensors and analysis packages, it is becoming easier to collect large amounts of data in the field—automated tagging of metadata will be essential, along with data quality checks. As multiple digital libraries come into existence, there is also the need to be able to navigate and relate provenance across such entities. As previously mentioned, there is a need to search across multiple, distributed catalogs, and maintaining persistence and consistency across such catalogs is an open issue. Warner et al. address interoperability issues that arise in this context.

The article by Zimmerman et al. highlights the issue of not only how scientists are currently trained to look for authoritative knowledge, but also where trust currently lies in the system. The challenge of online digital library environments is to create such trusted sources. Google search is "trusted" not because it is a curated source, but just by the sheer volume of data that it indexes and because it provides credible results most of the time. Digital libraries, on the other hand, will be trusted because they are *well-curated*, which will likely be judged on the basis of the science results that are produced by using resources (data and tools) from such libraries. Librarians and archivists, whose task it is to carefully index, classify, and organize information for access and use are understandably overwhelmed at the daunting task of organizing and curating the continuously flowing deluge of data and derived products. The support for user annotations and systems documenting the provenance of data and derived products, discussed in some papers in this issue, will be needed in such systems to guide future users.

The international dimension brings other issues related to data sharing. In addition to different cultures of sharing in different science disciplines, there is now the culture of sharing across multiple countries and multiple regulatory domains. Physics and astronomy projects have developed vibrant international collaborations, e.g., the Large Hadron Collider at CERN and the International Virtual Observatory Alliance. Expensive instruments may need to be located at a particular place on the planet, but a worldwide community of scientists shares the data streams generated by these instruments. However, the sharing of data in disciplines where the information has other uses and implications—e.g., natural resources, national security, global climate change, culture and history—may be limited by regulatory or economic policies. The OneGeology project has recently been launched in order to "map the geology of the planet" (http://www.onegeology.org). The project has gained rapid momentum, indicating a latent desire among the geologic community across countries to share geologic information, even if it is at a lower resolution.

An important aspect to consider is the structure of the science community in a given country. The GEON project has international collaborations and has conducted workshops in India, China, and Russia. In all cases, the collaborators are from (federal government-run) universities or federal research labs. Also, much of the scientific data in these countries is held by federal agencies and not easily sharable. In more than one case, we have been informed that recent, higher resolution data sets are available, but can be shared only in person, i.e., if a research collaborator visits the institutions in person. There is no possibility, as of now, of making such data available online.

It is quite understandable that countries that are newly emerging on the global economic and scientific arena are trying to be careful in how they share valuable assets such as their scientific data. The goal should be to establish trusted collaborations along with appropriate information technology solutions that will allow controlled sharing of valuable information. For example, this is being done in the GEON project by establishing a GEON Portal in India, designed to share data within India and make that data available via the main GEON Portal in the US. Establishing the goal of enabling local sharing first, followed by the equally important goal of sharing internationally, was necessary to make progress. While there are many large, multi-institutional collaborative projects in the US, it has been our experience that such collaborations are rare in many other countries. In particular, we do not see many instances of collaboration between Computer Science and other science disciplines, as one sees in the United States. International collaborations in e-science can serve to demonstrate to science communities in other countries how such collaborations between Computer Science, Library and Information Science, and various science and humanities disciplines can have a transformative effect on the research and educational enterprise.

There are, undoubtedly, many challenging issues to consider. In the long run, will the system tolerate asymmetries in sharing (e.g., across countries) and, if so, under what circumstances? How does one establish the veracity of shared data and develop trust relationships across international digital libraries? What agreed upon protocols are needed? What happens if one party wishes to unilaterally withdraw all its data?

The era of e-science and cyberinfrastructure is just beginning and everyone across the globe should be ready for an exponential takeoff!