OPEN ACCESS TO DIGITAL INFORMATION: OPPORTUNITIES AND CHALLENGES IDENTIFIED DURING THE ELECTRONIC GEOPHYSICAL YEAR

WK Peterson

Laboratory for Atmospheric and Space Physics, University of Colorado, Innovation Drive, Boulder, Colorado 80303 USA
Email: Bill.Peterson@lasp.colorado.edu

ABSTRACT

The vision of the Electronic Geophysical Year (eGY) is that we can achieve a major step forward in geoscience capability, knowledge, and usage throughout the world for the benefit of humanity by accelerating the adoption of modern and visionary practices such as virtual observatories for managing and sharing data and information. eGY has found that the biggest challenges to implementing the vision are educating program mangers and senior scientists on the need for modern data management techniques and providing incentives for practitioners of the new field of geoinformatics.

Keywords: Virtual observatory, Data management, Data archiving, Data policy, Informatics, Geoscience

1 INTRODUCTION

The Electronic Geophysical Year (*e*GY) is an effort of the international geosciences community to recall the significant advances of the International Geophysical Year (IGY) fifty years ago and to remind ourselves of the major advances that were facilitated by sharing the accumulated IGY data. Rather than focus on events of fifty years ago, *e*GY chose to focus on data stewardship by addressing the issues of open access to data, data preservation, data discovery, data rescue, capacity building, and outreach. In particular *e*GY has focused on the development of virtual observatories (Fox, 2008). Developing and implementing virtual observatories involves all aspects of the evolving new discipline of informatics, which bridges the gap between geoscience and computer science (Baker et al., 2008). This new field has been called geoinformatics.

Virtual observatories require more than open access to data to operate. A virtual observatory must have access to a detailed, computer readable, catalog of data holdings, which are accessible to computer generated queries. Language-independent interfaces to databases (e.g. Application Programming Interfaces or API's) are the basis of virtual observatories. Documenting databases and implementing API's to support virtual observatories are new and expensive requirements for data providers. *e*GY has found that the biggest challenges to implementing and providing data to virtual observatories are educating program managers and senior scientists on the need for modern data management tools and practices and providing adequate incentives for practitioners of the new field of geoinformatics.

In this article we discuss both the major opportunities and challenges to open access to digital information identified during the Electronic Geophysical Year.

2 OPPORTUNITIES

The opportunities identified in implementing the eGY vision include:

- Worldwide access to extremely large and distributed databases allows for developing more comprehensive views of our dynamic geophysical environment.
- The direct machine-to-machine communications implemented in virtual observatories makes geoscientists more productive by allowing them to focus on the content of the data, not details of accessing and formatting.
- Geoscientific data can be updated in real time, providing insights into geodynamical variations over all time scales.
- Virtual observatories make data available to a diverse set of new users unfamiliar with details and locations of geophysical data and information.

Rather than directly addressing all the opportunities recent advances in internet technology have facilitated, we focus here on the opportunities virtual observatories provide by making data and information available to new classes of users.

Figure 1 shows that web services are a relatively new feature on the internet. They are a logical extension of the internet protocols that were designed to provide web page content directly to a user's web browser using markup languages such as HTML. One response to rapid growth of web content was the advent of search engines. Another, and perhaps more important response, was the development of standards such as XML and application programming interfaces (API) for web services that enable direct machine to machine transfer of web content.

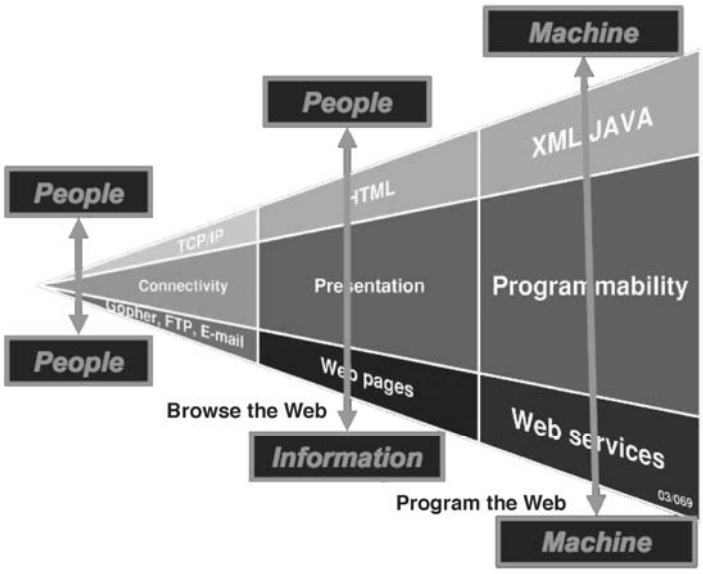

Figure 1. Schematic view of the development of the internet data and information transfer illustrating the history of data and information transfer methods. Adapted from a web page of the now defunct Distributed Systems Technology Center (DSTC) that was supported until June 2006 by the Australian Government's Cooperative Research Center program.

The flexibility of these new web protocols and standards are now being exploited by developers of virtual observatories to provide machine-to-machine transfer of data and information over the internet. As discussed by Fox (2008) virtual observatories require development of a common vocabulary describing the data and information being transferred. In the space physics community users are migrating to the Space Physics Archive Search and Extract (SPASE) Query Language (Narock et al., 2008). Controlled vocabularies such as SPASE are discipline specific. Only experts in the field can understand the relationship between terms in the discipline specific controlled vocabulary and fully exploit data described. As noted by Fox et al. (2007), application of semantic web methods and technologies (see for example, Berners-Lee & Miller, 2002) is an effective method to address the challenging

problem of integrating data and controlling vocabularies from diverse disciplines, thus making them accessible to non-experts.

As noted by CoBabe-Ammann et al. (2007) "Scientists aren't the only ones looking for data. From policy makers faced with natural hazards to teachers looking to bring data into the classroom to the "citizen scientists" searching for more information on a topic near and dear to their heart, people without specialized scientific knowledge are coming to the Internet looking for data. What they find isn't always an easy pathway to knowledge." Developing web services to serve these non-specialists provides an opportunity to apply state-of-the-art semantic web technology to current and relevant problems.

Identifying, addressing, and meeting the needs of these new classes of non-expert users is the greatest opportunity provided by virtual observatory technology.

3 CHALLENGES

There are many challenges to providing free and open access to geoscientific data. The greatest challenge is identifying what needs to be archived and obtaining continuing resources to support growing archives. For digital data there are the added complications associated with computer security, commercial and/or governmental restrictions and requirements, and the rapid evolution of hardware and software. During the Electronic Geophysical Year (*e*GY) we have identified three new challenges arising from the use of web services in virtual observatories and the subsequent requirement for documenting on-line data sets using controlled vocabularies or formal ontologies. These challenges are described below.

3.1 The archiving challenge

In the IGY era the challenge was to transmit data across the "iron curtain" dividing scientists and their data and the information derived from them. During IGY data archiving and distribution were addressed by the formation of World Data Centers and associated World Data Center Panel under the auspices of the International Council for Science (ICSU). These organizations served the geoscientific community well. The "iron curtain" has since fallen. The internet has revolutionized data distribution. Computer technology has greatly expanded the volume and types of data sets that are exchanged and need to be archived.

In the *e*GY era web sites and web services are dynamic; they are not archives. More importantly the migration to web data services by the world data centers has not been uniform. *e*GY participants have joined in discussions with various scientific bodies about the processes involved in identifying, capturing, and archiving selected geophysical data sets. In particular *e*GY was represented on the Strategic Committee on Information and Data (SCID) organized by ICSU. The SCID report was discussed at the fall 2008 ICSU meeting and a committee formed to begin implementing its recommendations. At the time of this writing, it is not clear how the SCID recommended ICSU World Data System would be implemented. A copy of the SCID report is available on the ICSU web site at http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/2123_DD_FILE_SCID_Report.pdf

3.2 The educational challenge

Educating scientists and science managers about modern data management techniques was identified as the greatest challenge to free and open access to data in the *e*GY era.

Data exchange between geographically distributed investigators enabled by the first and second generations of internet data distribution systems provided a rich environment for scientific advances. These data exchanges were developed from and are still based on personal interactions between geographically distributed investigators. Data exchanged in first and second-generation systems are stored in a variety of formats with patchy and non-uniform documentation. Most importantly there are no systematic ways to introduce new or significantly expanded or modified data sets. This has let to mostly informal discipline specific data sets. Investigators who are not personally introduced and are able to interact with to one of the users of these disciple specific distributed data sets cannot effectively use the data. The vision of eGY is to enable these new users to have effective access to all relevant data

through the third generation internet technology illustrated in Figure 1. As noted in section 2 virtual observatory technology opens discipline specific data sets not only to new investigators in closely related disciplines but to the broader community that includes policy makers, educators, and citizen scientists.

In the course of advocating virtual observatory technology eGY participants often were confronted by working scientists and more often by their managers questioning the need to upgrade their data systems to address the documentation requirements of virtual observatories, described in section 3.3, when the existing set of web pages and ftp sites were serving their communities adequately. There is always resistance to new technologies; there are still a few senior managers who do not use a personal computer. Resistance to new technology is overcome by use, education, and attrition.

*e*GY participants at two annual meetings focused on the need for educating current and future geoscientists on modern data management practices, including virtual observatory technology. eGY believes that the increasing volume and complexity of data needed for research requite that investigators be aware of the powerful tools now available to organize, distribute, and archive data. eGY advocates that a course in data management be required of all new investigators. In particular *e*GY is looking for well organized and jargon free materials describing virtual observatory technology to current investigators. The field, however, is so new that these materials have not yet been developed. *e*GY participants are currently active in the American Geophysical Union (AGU) Earth and Space Science Informatics (ESSI) focus group, and the other similar groups in the Geological Society of America (GSA), the International Union of Geodesy and Geophysics (IUGG), and the European Geophysical Union (EGU) to build the critical mass to effectively address this challenge.

3.3 The documentation challenge

Virtual observatory technology allows identification, transfer, and re-formatting of data from any computer on the internet. Data are presented to an investigator on a computer on his desk where he can manipulate and process them independently of the data provider. The technology is based on standards that use a controlled vocabulary or complete ontology describing the data in sufficient detail. There are also standards for identifying the data so that they can be found from search engines.

Documentation of data sets for access by virtual observatories is a new and time intensive requirement for data providers. To take full advantage of virtual observatory technology, someone very familiar with the data, the controlled vocabulary or ontology, and virtual observatory technology must do data set architecture design and documentation. Data set design should be done early in any new investigation so that all of the information necessary for documentation are identified and collected. Persons best suited to this task are either scientists with a strong interest in data management or data professionals who are very familiar with the data set of interest. These individuals are practicing the new science/engineering field of informatics. Baker et al. (2008) have defined and described this new field. Only a small subset of geoscientific investigators currently has the technical expertise in informatics. They are not yet able to efficiently and effectively design, document, and implement data sets for use by virtual observatories.

The reason there are so few qualified geoscientists or data professionals who also are familiar with geoinformatics is because understanding the tools of modern data management has not been a priority of managers of geoscientific projects. This includes deans, department heads, and in some cases managers of national programs. The few people who have pioneered the development of informatics have, in general, not been rewarded with increased status and salary as have their peers who have focused on the publishing papers based on the data or developing applications that use existing data sets. *e*GY participants at two annual meetings noted this disparity and recommended that education in modern data management principles be required of all geoscientists as described in Section 3.2.

*e*GY participants also suggested that the visibility and status of informatics be increased by establishing fora and publications for the discussion and dissemination of information and results on the subject. During the *e*GY the American Geophysical Union (AGU), the European Geophysical Union (EGU), the International Union of Geodesy and Geophysics (IUGG), and the Geological Society of America (GSA) have formed groups to organize discussions of informatics at their annual meetings. In addition the journal of Earth Science Informatics (Fox, 2008) has been created for the publication of geoinformatics material.

4 CONCLUSION

Digital information can flow freely and be available for use without having to pay attention to details such as location and format if 1) we continue to inform and educate new and active geoscientists of modern data management principles and tools; and 2) we develop clear paths for advancement for informatics professionals. The development of virtual observatories to meet the needs of non-expert users provides a challenging and potentially rewarding opportunity to practitioners of the new science of geoinformatics.

5 ACKNOWLEDGEMENTS

The author thanks the many *e*GY participants who contributed to his education in geoinformatics during the *e*GY. He would especially like to thank Emily CoBabe-Ammann, Peter Fox, Mark Parsons, and Charles Barton.

6 REFERENCES

Baker, D.N., Barton, C.E., Peterson, W.K., and Fox, P. (2008) Informatics and the 2007–2008 Electronic Geophysical Year. *EOS, Trans. Amer. Geophys. U. 48(25*), 485-487

Berners-Lee, T., and Miller, E. (2002) The semantic web lifts off. *The European Research Consortium for Informatics and Mathematics (ERCIM) News No. 51.*

CoBabe-Ammann, E., Peterson, W.K., Baker, D., Fox, P., and Barton, C. (2007) The Electronic Geophysical Year (2007-2008): eScience for the 21st century. *The Leading Edge 26(10*), 1294-1295

Fox, P., McGuinness, D., Raskin, R., and Sinha, K. (2007) A volcano erupts: semantically mediated integration of heterogeneous volcanic and atmospheric data. Proceedings of the *ACM first workshop on CyberInfrastructure: information management in eScience*, Lisbon, Portugal. Retrieved from the WWW, January 21, 2010: http://doi.acm.org/10.1145/1317353.1317355

Fox, P. (2008) Virtual observatories in geosciences. *Earth Science Informatics 1(1*), 3-4. Retrieved from the WWW, January 21, 2010: http://www.springerlink.com/content/0322621781338n85/

Narock, T.W. and King, T. (2008) Developing a SPASE Query Language. *Earth Science Informatics 1(1*), 43-48. Retrieved from the WWW, January 21, 2010: http://www.springerlink.com/content/f4h25211t7772u33/