Towards Secure Cyberinfrastructure for Sharing Border Information

Ann Gates[1], Vladik Kreinovich[1], Luc Longpré[1],
Paulo Pinheiro da Silva[1], G. Randy Keller[2]
[1]Department of Computer Science
[2]Department of Geological Sciences
University of Texas at El Paso
500 W. University, El Paso, TX 79968, USA
agates@utep.edu, vladik@utep.edu, longpre@utep.edu,
paulo@utep.edu, keller@utep.edu

Abstract

In many border-related issues ranging from economic collaboration to border security, it is extremely important that bordering countries share information. One reason why such sharing is difficult is that different countries use different information formats and data structures. It is therefore desirable to design infrastructure to facilitate this information sharing.

UTEP is a lead institution in a similar NSF-sponsored multi-million geoinformatics project, whose goal is to combine diverse and complex geophysical and geographical data stored in different formats and data structures. We describe our experience in using and developing related web service techniques, and we explain how this experience can be applied to border collaboration.

Since many security-related data are sensitive, we describe how to make sure that the designed cyberinfrastructure provides secure sharing.

1 Practical Problem: Need to Combine Geographically Separate Computational Resources

In different knowledge domains in science and engineering, there is a large amount of data stored in different locations, and there are many software tools for processing this data, also implemented at different locations. Users may be interested in different information about this domain.

- Sometimes, the information required by the user is already stored in *one of* the *databases*.

For example, if we want to know the geological structure of a certain region in Texas, we can get his information from the geological map stored in Austin. In this case, all we need to do to get an appropriate response to the query is to get this data from the corresponding database.

- In other cases, different pieces of the information requested by the user are *stored at different locations*.

  For example, if we are interested in the geological structure of the Rio Grande Region, then we need to combine data from the geological maps of Texas, New Mexico, and the Mexican state of Chihuahua. In such situations, a correct response to the user's query requires that we access these pieces of information from different databases located at different geographic notations.

- In many other situations, the appropriate answer to the user's request requires that we not only collect the relevant data, but that we also use some *data processing* algorithms to process this data.

  For example, if we are interested in the large-scale geological structure of a geographical region, we may also use the gravity measurements from the gravity databases. For that, we need special algorithms to transform the values of gravity at different locations into a map that describes how the density changes with location. The corresponding data processing programs often require a lot of computational resources; as a result, many such programs reside on computers located at supercomputer centers, i.e., on computers which are physically separated from the places where the data is stored.

The need to combine computational resources (data and programs) located at different geographic locations seriously complicates research.

2 Centralization of Computational Resources – Traditional Approach to Combining Computational Resources; Its Advantages and Limitations

Traditionally, a widely used way to make these computational resources more accessible was to move all these resources to a *central location*. For example, in the geosciences, the US Geological Survey (USGS) was trying to become a central depository of all relevant geophysical data. However, this centralization requires a large amount of efforts: data are presented in different formats, the existing programs use specific formats, etc. To make the central data depository efficient, it is necessary:

- to reformat all the data,
- to rewrite all the data processing programs – so that they become fully compatible with the selected formats and with each other,
- etc.

The amount of work that is needed for this reformatting and rewriting is so large that none of these central depositories really succeeded in becoming an easy-to-use centralized database.

3 Cyberinfrastructure – A More Efficient Approach to Combining Computational Resources

Cyberinfrastructure technique is a new approach that provides the users with the efficient way to submit requests without worrying about the geographic locations of different computational resources – and at the same time avoid centralization with its excessive workloads. The main idea behind this approach is that *we keep all (or at least most) the computational resources*

- *at their current locations*,
- *in their current formats*.

To expedite the use of these resources:

- we supplement the local computational resources with the "metadata", i.e., with the information about the formats, algorithms, etc.,

- we "wrap up" the programs and databases with auxiliary programs that provide data compatibility into *web services*,

and, in general, we provide a cyberinfrastructure that uses the metadata to automatically combine different computational resources.

For example, if a user is interested in using the gravity data to uncover the geological structure of the Rio Grande region, then the system should automatically:

- get the gravity data from the UTEP and USGS gravity databases,
- convert them to a single format (if necessary),
- forward this data to the program located at San Diego Supercomputer Center, and
- move the results back to the user.

This example is exactly what we are designing under the NSF-sponsored Cyberinfrastructure for the Geosciences (GEON) project; this is similar to what other cyberinfrastructure projects are trying to achieve.

4 What Is Cyberinfrastructure: The Official NSF Definition

According to the final report of the National Science Foundation (NSF) Blue Ribbon Advisory Panel on Cyberinfrastructure, "a new age has dawned in scientific and engineering research, pushed by continuing progress in computing, information, and communication technology, and pulled by the expanding complexity, scope, and scale of today's challenges. The capacity of this technology has crossed thresholds that now make possible a comprehensive 'cyberinfrastructure' on which to build new types of scientific and engineering knowledge environments and organizations and to pursue research in new ways and with increased efficacy.

Such environments and organizations, enabled by cyberinfrastructure, are increasingly required to address national and global priorities, such as understanding global climate change, protecting our natural environment, applying genomics-proteomics to human health, maintaining national security, mastering the world of nanotechnology, and predicting and protecting against natural and human disasters, as well as to address some of our most

fundamental intellectual questions such as the formation of the universe and the fundamental character of matter."

5 Geoinformatics: Cyberinfrastructure for the Geosciences

Geoinformatics is a term that appears to have been independently coined by several groups around the world to describe a variety of efforts to promote collaboration between computer science and the geosciences to solve complex scientific questions. Fostered by the leadership within the National Science Foundation, Geoinformatics has emerged as an initiative within the Earth Sciences Division to address the growing recognition that the Earth functions as a complex system and that existing information science infrastructure and practice used by the geoscience community are inadequate to address the many difficult problems posed by this system. In addition, there is now widespread recognition that successfully addressing these problems requires integrative and innovative approaches to analyzing, modeling, and developing extensive and diverse data sets. Currently,

- the chaotic distribution of available data sets,
- lack of documentation about them, and
- lack of easy-to-use access tools and computer modeling and analysis codes

are major obstacles for scientists and educators alike. These obstacles have hindered scientists and educators in the access and full use of available data and information derived from it, and hence have limited scientific productivity and the quality of education. However, recent advances in fields such as

- computational methods,
- visualization, and
- database interoperability

provide practical means to overcome such problems.

Earth Science is a discipline that is strongly data driven, and large varied data sets are often developed by researchers and government agencies. However, the geosciences community knows all too well the difference between a large data set and a useable database. To be

sure, a number of databases exist, but almost none of them are truly complete, error free as is practical, easily accessible, and simple to use. The ultimate vision of the Geoinformatics initiative is a highly interconnected data system populated with high quality, freely available data, as well as, a robust set of software for analysis, visualization, and modeling. This system would feature rich and deep databases and convenient access.

The development of the capability to construct, organize, and verify an Earth Science data system is a natural, and indeed essential, step for the Earth Sciences to move forward so that we can understand the Earth as a system and meet societal needs. Most Earth Science problems are inherently 4-D (x, y, z, t) in nature involving the subsurface and variation with time. Thus, their solution requires data analysis that is far more complex than provided by traditional geographic information systems (GIS). The extent, complexity, and sometimes primitive form of existing data sets and databases, as well as the need for the optimization of the collection of new data, dictate that only a large, cooperative, well coordinated, and sustained effort will allow the community to attain its scientific goals. With a strong emphasis on ease of access and use, the resulting data system would be a very powerful scientific tool to reveal new relationships in space and time and would be an important resource for students, teachers, the public at large, governmental agencies and industry.

Cyberinfrastructure is a new term that refers to the information technology infrastructure that is needed to:

- manage, preserve, and efficiently access the vast amounts of Earth Science data that exist now and the vast data flows that will be coming online as projects such as EarthScope (www.earthscope.org) get underway;
- foster integrated scientific studies that are required to address the increasingly complex scientific problems that face our scientific community;
- accelerate the pace of scientific discovery and facilitate innovation;
- create an environment in which data and software developed with public funds are preserved and made available in a timely fashion; and
- provide easy access to high-end computational power, visualization and open source software to researchers and students.

The task of creating a cyberinfrastructure for the Earth Sciences is daunting due to the large volume and diversity of the data, as well as, the extreme differences in data formats, storage and computing systems, as well as differing conventions, terminologies, and ontological frameworks across disciplines. One way to think about this is that the ultimate goal is provide one with the tools and data needed to do better, more creative science by minimizing the effort needed to look for data, research the background of a topic, and make software run properly. Another consideration is that when data and information are entered into an organized system, they can be easily found and unexpected relationships can be discovered via queries in "Google-like" fashion. An earth scientist should think in terms of discovering many of the relationships between phenomena that led to plate tectonics in days instead of years.

6 GEON Project

Building on the geoscience/computer science partnership fostered by our early database construction and data dissemination efforts, our group has played an active role in the Geoinformatics initiative since its early days. In collaboration with a number of colleagues, we have been awarded several related NSF grants:

- the Southwest GeoNet;
- a small Information Technology Research (ITR) project that is a partnership between the University of Texas at El Paso and Arizona State University, and
- a large ITR grant funds the GEON (GEOscience Network) project.

The focus of GEON is the pressing need in the geosciences to:

- craft the many relatively raw data sets in the Earth Science community into mature databases that can grow and evolve;
- interlink and share these multidisciplinary databases;
- create a robust toolbox of open source software for analysis, modeling, and visualization; and

- provide the information technology infrastructure to manage and explore a highly distributed and diverse network.

GEON is a true partnership between Computer Science and Earth Science researchers, and the scientific goal of this project is to facilitate efforts to understand complex problems focusing on the 4-D structure and evolution of continents. To rise to this challenge, we formed a coalition of researchers with key Computer Science expertise and researchers representing a broad cross-section of Earth Science sub-disciplines.

The creation of GEON is a first step in developing the critical cyberinfrastructure necessary to achieve the vision of Geoinformatics and facilitate other research initiatives, in particular EarthScope. GEON is working closely with organizations such as IRIS, the U.S. Geological Survey, SCEC, and UNAVCO as well as other IT efforts within the Earth Science community. In particular, the U.S. Geological Survey has joined as a major partner and has made creation of key GEON-related databases a priority effort over the next several years.

7 Cyberinfrastructure for the Geosciences: Challenges

Creating the GEON cyberinfrastructure to integrate, analyze, and model 4D data poses fundamental IT research challenges due to:

- the extreme heterogeneity of geoscience data formats, storage and computing systems and, most importantly,
- the ubiquity of "hidden semantics" and differing conventions, terminologies, and ontological frameworks across disciplines.

As the prototype for a national cyberinfrastructure in the geosciences, our guiding principle is embracing heterogeneity at all levels-hardware, software, networking, as well as information structures.

Creating this infrastructure involves basic as well as applied research in Information Technology and use of state-of-the-art information technologies. GEON IT research focuses on:

- modeling,
- indexing,

- semantic mediation,
- visualization of multi-scale 4D data, and
- creation of a prototype GEON Grid.

An important contribution is embarking on the definition of a Unified Geosciences Language System (UGLS), to enable semantic interoperability. For example, a stratigraphic layer of rock often changes names across state lines, and think of all the possible answers to a "Google-like" query about the age of a rock unit or the magnitude of an earthquake and of all the places where one might find this information.

We have started creating a portal to provide access to the GEON environment, which will include:

- advanced query interfaces to distributed, semantically-integrated databases,
- Web-enabled access to shared tools, and
- seamless access to distributed computational, storage, and visualization resources and data archives.

Various GEON-like grid efforts, such as GriPhyN, NEESGrid, NBII, and BIRN, all indicate the readiness of the Computer Science community to provide the necessary interoperable infrastructure, and testify to the value of integration of Computer Science with major science and education initiatives.

In order to ensure that the scope of GEON was manageable, linkage and refinement of existing and emerging databases is being emphasized, and two testbeds:

- the mid-Atlantic and
- the Rocky Mountain region

were identified to focus the GEON geoscience research effort geographically. These regions were selected:

- due to the variety of geological issues embodied within them that require integration of multi-disciplinary database, and
- because they are areas of expertise for the GEON geoscience research team.

The ultimate goal of GEON research is to significantly impact large multi-scale geoscience research programs such as Earthscope, as well as individuals and smaller groups of researchers with the goal of facilitating the development of a culture in which data are shared, archived and rapidly disseminated across the Earth sciences, much like IRIS has done within seismology. Many disciplinary geoscience database projects are already underway, indicating the readiness of the community to participate in such a national-scale effort. By facilitating the use of large and diverse data sets, we believe that the scientific community will make major scientific discoveries and create new and exciting scientific paradigms that lead is into the post-plate tectonics era.

In the case of both test beds, our goal is to pursue an ambitious research agenda in Earth Science stressing integrated studies while working with Computer Science colleagues to create a cyberinfrastructure that pushes the envelop in their field. We hope to create an environment in which researchers will be able to see benefits for their personal efforts and will want to contribute data, software, and ideas.

8 GEON's Objective: Brief Summary

The goal of GEON is:

- to advance the field of geoinformatics,
- to prepare and train current and future generations of geoscience researchers, educators, and practitioners in the use of cyberinfrastructure to further their research, education, and professional goals.

Geoinformatics will foster:

- new interdisciplinary research, for example, the gravity modeling of 3D geological features, such as plutons;
- study of active tectonics by integrating LiDAR data and geodynamics models; and
- study of lithospheric structure and properties across diverse tectonic environments.

GEON is based on a service-oriented architecture (SOA) with support for

- "intelligent" search,

- semantic data integration,
- visualization of 4D scientific datasets, and
- access to high performance computing platforms for data analysis and model execution – via the GEON Portal.

9 Our Experience Can Be Applied to Border Collaboration

While focused on Earth Sciences, GEON cyberinfrastructure is generic and broadly applicable to a variety of other sciences and other application domains.

One possible application domain is border collaboration. In many border-related issues ranging from economic collaboration to border security, it is extremely important that bordering countries share information. One reason why such sharing is difficult is that different countries use different information formats and data structures. It is therefore desirable to design infrastructure to facilitate this information sharing.

In principle, it is possible to transform all this data into a single format and store it into a single large database. However, in practice, this is rarely a feasible solution, for two reasons:

- first, transforming the large amounts of data into different formats would require a large amount of effort;
- second, the agreement on a single format for storing common data may be politically difficult.

Our experience in designing cyberinfrastructure for the geosciences shows that with a much smaller amount of effort, and with no need to make complex political decisions, it is possible to combine different border-related databases by an appropriate cyberinfrastructure, by "wrapping" up each computational resource into the corresponding web service.

10 How to Make Cyberinfrastructure Secure

Since many security-related border-related data are sensitive, it is important to make sure that the designed cyberinfrastructure provides *secure* sharing.

Traditionally, this security problem is solved only on the level of *data sharing*. The problem of secure data sharing and transmission is successfully solved, e.g., by using different encryption schemes such as the RSA algorithm (see, e.g., [7]). Data exchange that uses such an encryption is successfully implemented in different e-commerce applications such as amazon.com.

In general cyberinfrastructure applications, the situation is more complex: it is not enough to protect the data during transmission. In order to process data, we must decode it, and this is where the data becomes vulnerable. So, if we use remote computational resources to process data, we must make sure that this data is not compromised by the computations, i.e., that, if necessary, the remote performs the computations without being able to learn the underlying data.

There exist various protocols for computing without learning the underlying data. Most of these protocols are based on a protocol for privacy-preserving auctions proposed by Naor, Pinkas, and Sumner in [10]. It is therefore desirable to extent this protocol to security and privacy protection on web services.

In [6, 9], we show that a protocol from [10] can indeed be modified to preserve privacy of customers using web services. As a case study, we considered a commercially important problem of on-line tax preparation, where a customer wants to use an on-line tax computation tool, but:

- a customer wants to keep his or her financial data private, and
- the remote tax computation site wants to keep their software private, so that the user will not be able to copy it and use it for free.

Without remote access, it is difficult to achieve such privacy:

- if we move all the customer's data to the remote site, then the customer cannot be sure than his or her data will be protected;
- on the other hand, if we download a copy of the tax computation program to the customer site, the program owner cannot be sure that the program is protected from copying.

It turns out that with remote access, it is possible to only partially move the data and the program and thus, guarantee privacy for both sides.

A similar problem appears when we want to compute overall characteristics of a border area based on information that both sides do not necessarily want to share in detail:

- due to security concerns or
- due to the need to keep commercial secrets.

Of course, to ensure privacy, we must perform additional computations – e.g., encoding and decoding require additional computation time. We therefore recommend that this protocol be used only for those web services for which privacy preservation is more important than efficiency.

11 Conclusions

In many border-related issues ranging from economic collaboration to border security, it is extremely important that bordering countries share information. One reason why such sharing is difficult is that different countries use different information formats and data structures. It is therefore desirable to design infrastructure to facilitate this information sharing.

UTEP is a lead institution in a similar NSF-sponsored multi-million geoinformatics project, whose goal is to combine diverse and complex geophysical and geographical data stored in different formats and data structures. We show how our experience in using and developing related web service techniques can be applied to border collaboration. We describe how to make sure that the designed cyberinfrastructure provides secure sharing.

Acknowledgments

This work was supported in part by NASA under cooperative agreement NCC5-209, by NSF grants EAR-0225670 and DMS-0532645, Army Research Lab grant DATM-05-02-C-0046, Star Award from the University of Texas System, and Texas Department of Transportation grant No. 0-5453.

References

[1] M. S. Aguiar, G. P. Dimuro, A. C. R. Costa, R. K. S. Silva, F. A. Costa, and V. Kreinovich, "The Multi-Layered Interval Categorizer Tesselation-Based Model", In:

C. Iochpe and G. Câmara (eds.), *IFIP WG2.6 Proceedings of the 6th Brazilian Symposium on Geoinformatics Geoinfo'2004*, Campos do Jordão, Brazil, November 22–24, 2004, pp. 437–454. ISBN 3901882200.

[2] R. Aldouri, G. R. Keller, A. Gates, J. Rasillo, L. Salayandia, V. Kreinovich, J. Seeley, P. Taylor, and S. Holloway, "GEON: Geophysical data add the 3rd dimension in geospatial studies", *Proceedings of the ESRI International User Conference 2004*, San Diego, California, August 9–13, 2004, Paper 1898.

[3] M. G. Averill, K. C. Miller, G. R. Keller, V. Kreinovich, R. Araiza, and S. A. Starks, "Using Expert Knowledge in Solving the Seismic Inverse Problem", *Proceedings of the 24nd International Conference of the North American Fuzzy Information Processing Society NAFIPS'2005*, Ann Arbor, Michigan, June 22–25, 2005, pp. 310–314.

[4] M. Ceberio, S. Ferson, V. Kreinovich, S. Chopra, G. Xiang, A. Murguia, and J. Santillan, "How to take into account dependence between the inputs: from interval computations to constraint-related set computations, with potential applications to nuclear safety, bio- and geosciences", *Proceedings of the Second International Workshop on Reliable Engineering Computing*, Savannah, Georgia, February 22–24, 2006, pp. 127–154.

[5] M. Ceberio, V. Kreinovich, S. Chopra, and B. Ludäscher, "Taylor Model-Type Techniques for Handling Uncertainty in Expert Systems, with Potential Applications to Geoinformatics", *Proceedings of the 17th World Congress of the International Association for Mathematics and Computers in Simulation IMACS'2005*, Paris, France, July 11–15, 2005.

[6] W. Chen, *Two-Party Secure Computation of a Secret Boolean Function*, University of Texas at El Paso, Computer Science, Master's Thesis, December 2004.

[7] Th. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, MA, 2001.

[8] G. R. Keller, T. G. Hildenbrand, R. Kucks, M. Webring, A. Briesacher, K. Rujawitz, A. M. Hittleman, D. R. Roman, D. Winester, R. Aldouri, J. Seeley, J. Rasillo, R. Torres, W. J. Hinze, A. Gates, V. Kreinovich, and L. Salayandia, "A community effort to construct a gravity database for the United States and an associated Web portal", In:

A. K. Sinha (ed.), *Geoinformatics: Data to Knowledge*, Geological Society of America Publ., Boulder, Colorado, 2006, pp. 21–34.

[9] L. Longpré, V. Kreinovich, E. Freudenthal, M. Ceberio, F. Modave, N. Baijal, W. Chen, V. Chirayath, G. Xiang, and J. I. Vargas, "Privacy: Protecting, Processing, and Measuring Loss", *Abstracts of the 2005 South Central Information Security Symposium SCISS'05*, Austin, Texas, April 30, 2005, p. 2.

[10] M. Naor, B. Pinkas, and R. Sumner, "Privacy Preserving Auctions and Mechanism Design", *Proceedings of the 1st ACM Conference on Electronic Commerce*, Denver, Colorado, November 1999, pp. 129–139.

[11] E. Platon, K. Tupelly, V. Kreinovich, S. A. Starks, and K. Villaverde, "Exact Bounds for Interval and Fuzzy Functions Under Monotonicity Constraints, with Potential Applications to Biostratigraphy", *Proceedings of the 2005 IEEE International Conference on Fuzzy Systems FUZZ-IEEE'2005*, Reno, Nevada, May 22–25, 2005, pp. 891–896.

[12] C. G. Schiek, R. Araiza, J. M. Hurtado, A. A. Velasco, V. Kreinovich, and V. Sinyansky, "Images with Uncertainty: Efficient Algorithms for Shift, Rotation, Scaling, and Registration, and Their Applications to Geosciences", In: Mike Nachtegael, Dietrich Van der Weken, and Etienne E. Kerre (eds.), *Soft Computing in Image Processing: Recent Advances*, Springer Verlag (to appear).

[13] A. K. Sinha, *Geoinformatics: Data to Knowledge*, Geological Society of America Publ., Boulder, Colorado, 2006.

[14] R. Torres, G. R. Keller, V. Kreinovich, L. Longpré, and S. A. Starks, "Eliminating Duplicates Under Interval and Fuzzy Uncertainty: An Asymptotically Optimal Algorithm and Its Geospatial Applications", *Reliable Computing*, 2004, Vol. 10, No. 5, pp. 401–422.

[15] Q. Wen, A. Q. Gates, J. Beck, V. Kreinovich, and G. R. Keller, "Towards automatic detection of erroneous measurement results in a gravity database", *Proceedings of the 2001 IEEE Systems, Man, and Cybernetics Conference*, Tucson, Arizona, October 7–10, 2001, pp. 2170–2175.

[16] H. Xie, N. Hicks, G. R. Keller, H. Huang, and V. Kreinovich, "An IDL/ENVI implementation of the FFT based algorithm for automatic image registration", *Computers and Geosciences*, 2003, Vol. 29, No. 8, pp. 1045–1055.