Apr 22nd, 12:00 AM

Creating a culture of data integration and interoperability: librarians and Earth Science Faculty collaborate on a geoinformatics course

Michael Fosmire
Purdue University

Chris Miller
Purdue University

Creating a Culture of Data Integration and Interoperability: Librarians Collaborate on a Geoinformatics Course

C.C. Miller
Purdue University, U.S.A.
ccmiller@purdue.edu

Michael Fosmire
Purdue University, U.S.A.
fosmire@purdue.edu

Abstract

One aspect of the development of an e-infrastructure for research discovery that is often overlooked is the human element. Data is only as interoperable as the scientists who are willing to meet data standards and share their work with others. Information literacy in the data world has at least as many challenges as it does in the world of documents. For example, the provenance of data, who collected it, how they collected it, and whether and how it has been verified, are all very important factors for researchers to consider before incorporating external data into their analyses. With this in mind, the Purdue University Libraries and Earth and Atmospheric Sciences department partnered to offer a geoinformatics course to teach our next generation of scientists about the stores of data available for them to use and the power and limitations of networked tools and data structures to enhance the value and relevance of their own data, while fostering good data hygiene. Students will realize through hands-on activities why good data habits are critical to their professional success in the sciences. In this course, students follow a research project from the data gathering stage, through geospatial visualization and incorporation of external data sets, analysis of the combined data through workflow management software, comparison of results with data models, and, finally, curation of their data, culminating in its deposition in a disciplinary data repository, where it will be shared with and peer reviewed by classmates. The authors will describe how the goals and objectives of this course were determined and implemented and report findings from the pilot offering of this course.

e-science
instruction
geoinformatics

ENVIRONMENT AND LANDSCAPE

Some time between January 2003 and March 2007 a revolution occurred. In 2003, the National Science Foundation (NSF) released their "Revolutionizing Science and Engineering Through Cyberinfrastructure Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure" [Atkins, 2003] and led with a section titled "A Nascent Revolution" [Atkins, 2003, p. 9]. In some ways The Panel was referring to the same computer revolution most of the developed world is enjoying, but more specifically it is that current within the greater flood of more computers, more computation, and more automation in scientific research that brings with it previously impossible, vanguard methods of scientific inquiry. "New technology-mediated, distributed work environments are emerging to relax constraints of distance and time. These new research environments are linking together research teams, digital data and information libraries, high-performance computational services, scientific instruments, and arrays of sensors" [Atkins, 2003, p. 9]. In that trajectory of not only more computational power, not only more advanced systems, and not only more data, but more *access* to those resources, the Blue-Ribbon panel saw glimpses of a very cyber future in which, among other wholesale changes, "advanced computing is no longer restricted to a few research groups in a few fields...but pervades scientific and engineering research" [2003, p. 9] and in which "the classic two approaches to scientific research, theoretical/analytical and experimental/observational, have been extended to *in silico* simulation and modeling" [2003, p. 9]. In 2003, the origins of that future represented to the Panel "just the beginnings of a revolution" [2003, p. 10].

Well, the future is now. Or March 2007, rather, when NSF's Cyberinfrastructure Council published their "Cyberinfrastructure Vision for 21st Century Discovery" [2007]. In it, the NSF lays out a plan for directing and funding the initiatives designed to take the many and rapid advances in (mostly) large-scale computing architectures, communication and data transfer protocols and very large data storage capabilities and build from them a coordinated, integrated, interoperable cyberinfrastructure (CI) that, among other virtues, "serves as an agent for broadening participation and strengthening the nation's workforce in all areas of science and engineering" [The Council, 2007, p. 6]. Much of what NSF proposes is grid-level, that is, involving the supercomputing centers around the country and the TeraGrid that connects them at the research community. Yet they've taken care to acknowledge in their plan the fact that although e-infrastructure is an increasingly vital component of scientific research, these machines still require engineers, of sorts; scientists who can care for the data drawn from and fed into the electronic machines that produce or consume them. Just as much as high-end computing architectures must be relied upon to process and manipulate and share data, there is an equally important amount of human finesse that goes into the preparation, consumption, interpretation of and care for those data. The geoinformatics course described in this paper is an attempt by Library and cooperating Earth & Atmospheric Sciences faculty at Purdue University to address that niche within the emerging universe of cyberinfrastructure that will "both demand and support a new level of technical competence in the science and engineering workforce and in our citizenry at large" [The Council, 2007, p. 37].

So here we are again: barely re-energized from a fast and confusing decade in which librarians were spun dizzy and found clinging to a handful of retrenched values – information literacy among them – we are bracing for another, larger storm, the scale of which promises even greater challenges, both in magnitude and scale. Lest the horizon appear too bleak and foreboding for librarians trained at the kilobyte end of the spectrum or even weaned on the soft comfort of a paper catalog, there are two subcurrents within this greater rush toward a supercomputed future that provide more than a modicum of hope. Two little preservers bobbing between the swells.

The first of these is NSF's own unequivocal statement that "in the future, U.S. international leadership in science and engineering will increasingly depend upon our ability to leverage this reservoir of scientific data captured in digital form" [The Council, 2007, p. 22]. And as such, "ongoing attention must be paid to the education of the professionals who will support, deploy, develop, and design current and emerging cyberinfrastructure" [The Council, 2007, p. 38]. Sharp-eyed librarians will recognize the tenets of their own professional tradition, if not themselves, in that utopian scenario. Without promising the job to any one profession, The Council leaves ajar the door to a long, fruitful future for librarians and their computer science friends. Work is being done on this front already, of course, but there is still another

aspect of CI that librarians are finding to be ripe for the kind of attention a librarian is wont to pay an issue.

The development of *informatics disciplines (where '*' indicates any discipline one would care to mention), predating and contemporaneous to the march toward CI, has also enabled more and more opportunities for librarians, computer scientists, and like-minded domain scientists to recast the issues that either gird or inform CI (issues like data interoperability, semantic discovery and integration, long-term stewardship in decentralized systems) and attempt to wedge them into tight domain curricula. While bioinformatics is perhaps the best-known of all the *informatics branches, a younger, blossoming informatics, geoinformatics, is concerned with the geospatial elements of today's science. This means it is a remarkably fecund field for those interested in advancing both the systems and skills that will enable growing infrastructures to be even more capable of handling inherently and tangentially-geospatial datasets and applications. It is here, in this hotspot between CI and the semantic web and data futures and rapidly-developing, increasingly geospatially-savvy technologies, where The Libraries' and EAS' geoinformatics course intends to push on the "workforce" component of the NSF plan [The Council, 2007, ch. 5] and begin teaching our next generation of scientists about the stores of data available in online systems, the power and limitations of those networked tools and data structures, and the importance of good data hygiene.

The remainder of this paper will outline and annotate the design of the course, followed by a brief summary of where the course fits into the greater efforts of the NSF to build CI and then major and minor initiatives and developments with geoinformatics teaching and scholarship.

THE COURSE AT THE MEGABYTE LEVEL

While the hope is that the students will soon be advancing projects that draw at will from countless distributed super data stores and running simulations with cpu cycles leeched from centers on multiple continents, the course itself takes place in a more limited sphere. The authors would be remiss not to address the building blocks of CI in favor of the more advanced systems being developed by researchers in the field. As such, the course begins with units on basic computing environments, including Unix-based operating system environments and database structures and operations. Students are exposed to the relational model, SQL query and function syntax, and various manipulations of data.

Specifically geospatial flavors of data are then introduced, first as file-based raster and vector formats and then within the context of the geospatial database. Care is taken at each introduction to discuss primary uses of and sources for each of these data formats as well as the issues they raise for interoperability and storability.

Following the introduction to data formats, several weeks are spent introducing Geographic Information Systems (GIS) to students, first in terms of its foundational concepts (data layering and interaction, the divestment of display from data, and the great diversity of inputs) and then specifically as a tool for manipulation and analysis. Extra time is spent than would normally be in an introductory GIS unit on the place of GIS within the greater umbrella of geoinformatics. Although this means that the GIS-based exercises students are asked to perform are downsampled to a superficial sketch of real GIS work, the lessons in where geospatial data, data sharing, metadata, and techniques fit into the world of shared, interdisciplinary CI are the primary concern of the course.

The GIS unit segues into sections of the course designed to introduce data gathering (Global Positioning Systems [GPS] data), independent statistical procedures (review of statistics, MATLAB), and an unjustly slim section on the burgeoning world of inclusive scientific workflow tools; primarily the Kepler workflow program, but introduced by the (probably) more familiar, less wide-eyed Model Builder component of ESRI's ArcGIS suite.

As more resources and training become available in the world of geoinformatics, it is not unfathomable for the course to wade onto the grid and begin using distributed tools and systems in earnest. But the authors have day jobs, too, as working faculty librarians, and were concerned from the very start that a multi-discipline audience with varying backgrounds in GIS and e-tools and cyber-anything could easily become fractured and discombobulated when presented with too many, largely complex tools and operations.

PREPARING FOR THE UNKNOWN

In December 2007, the authors distributed an informal questionnaire to all students enrolled in the course. The questions were designed to gauge the level of familiarity and comfort with some of the technologies and resources and techniques that make up the course's foundational interpretation of geoinformatics. This interpretation includes the following concepts, each of which now represents a certain portion of the course syllabus but which, of course, taken together help the authors paint for students a fuller portrait: a landscape, perhaps, where connections *between* technologies and skills are more evident, and where portions of the research lifecycle with which students are likely to be unconcerned can be made more relevant.

Responses to the questionnaire revealed a rather surprising lack of previous experience with informatics and in some cases no experience at all with databases and even GIS. One resulting concern for the authors was that students were intending geoinformatics to be an introduction to GIS only. Although great care will be taken to encourage bigger-picture perspectives, it remains to be seen how students expectation and the realities of the broad, technologically-rich world of geoinformatics will or will not harmonize.

A GIMMICK TO EASE THE PAIN

As a means of making (at least the first half of) the experience more engaging, the authors chose to follow the model of MIT's *Environmental Detectives* and introduce a story arc that weaves through the first seven weeks of the course [Jenkins, Klopfer, Squire & Tan, 2003]. *Environmental Detectives* was an inaugural project within the MIT Teacher Education Program in association with The Education Arcade, also at MIT. Education Arcade games "reflect different pedagogical models, game genres, platforms, and classroom uses, showing the diverse ways in which educators of the future may be able to deploy computer and video games to enhance learning" [Jenkins *et al.* p. 2] and are built on "conceptual frameworks designed to support learning across math, science, engineering, and humanities curricula" [The Education Arcade, 2005] that were developed by the Microsoft-funded, joint MIT Comparative Media Studies Games-to-Teach Project.

The authors decided a faux scenario with elements of true data, true analysis, and feasible environmental repercussions would help usher students through the potentially gnarly introduction of GIS, databases, and data collection and analysis concepts. In our scenario, evidence of a water-borne contaminant is found in a river running adjacent to campus. Students are asked to collect well samples, along with various other relevant datasets such as surficial geology, soils, and elevation data, in an effort to build a model that will suggest to them potential origins of the substance. As students engage the mystery, and therefore engage the data and systems and analysis techniques required to collect and prepare the data in order to solve the puzzle, lessons about metadata, databases and data synthesis, GPS, and mostly GIS are introduced within, ideally, a compelling diegesis.

The mystery conceit lasts only through the first half of the course, however, by which time students will already have formulated and started projects of their own design. These semester projects will be planned and orchestrated as demonstrations of understanding of the core concepts in geoinformatics; the importance of metadata from beginning to end, the ability to discover, interpret, and consume disparate data sources and the systems that make them available, and the importance of a well-cured future for the products of their own research.

As these projects proceed, the rest of the lecture and lab modules move into the more integrated, interoperable world that stands as the template for the world of NSF CI. Data portals, both domain-specific and interdisciplinary, are introduced both as usable tools for student projects and as evidence of existing technologies germane to geoinformatics such as data markup, ontologies, and visualization.

Visualization is its own unit in the course, and the authors decided to embrace breadth rather than depth. In addition to sessions with various low-end visualization tools such as NASA WorldWind, Google Earth, and perhaps Yahoo! Maps or Google Maps, guest speakers will provide content that introduces more robust tools such as ArcGIS Explorer or Unidata's Integrated Data Viewer (IDV). Since NSF intends for great amounts of future research to be performed virtually and anticipates funding the programs that enable these new methods, it is increasingly important that students across domains are

familiar with the tools that their fields will be using to model past, present, and future scenarios.

Following the colorful and fantastic renderings of geospatial data in 2d, 3d, even 4d viewers, students are led to the climactic units of this librarian-led curriculum: data preservation, ontology, and metadata lectures and wrap-up labs. Students are encouraged to consider the processes and interfaces to which they were exposed during the course and ruminate about the role that metadata, data markup, semantic crosswalking and translation, and intelligent availability play in the lifecycle of data and data users. If the authors have done their job, these concepts will not be new and the importance of these issues will be foregone conclusions. Nonetheless, students will be led through the final (or initial, depending on one's perspective) steps of depositing data in an online repository.

CONCLUSION

The revolution to which NSF and others are responding by developing increasingly complex, but increasingly agile systems and services and technologies will only proceed and succeed if future generations of scientists are able to work upon that infrastructure. Infrastructure is only as good as the societies and cultures and development built thereupon. Likewise, future scientists are only as willing to work within the bothersome strictures necessitated by adherence to standards and interoperability if they've been trained and have been convinced of the benefits of doing so. Stated in analog one last time, even the once-fantastic technology of the book was lost on the illiterate, and the literacy skills required of scientists in the interdisciplinary, high-grade CI future proposed by NSF are perhaps more nascent than the revolution itself.

The geoinformatics course described herein will have finished its inaugural semester by April 2008. Lessons learned by students will ideally have prepared them to move further into their respective domains with eyes open to the opportunities that are guaranteed to exist for pushing disciplinary or interdisciplinary research into the powerful realm of big systems, data sharing, analysis, visualization and, of course, curation.

REFERENCES

The Education Arcade. (2005, October 15). About the education arcade. Retrieved December 9, 2007, from the World Wide Web: http://educationarcade.org/about

Atkins, D. E. (2003). Revolutionizing science and engineering through cyberinfrastructure report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Arlington, VA: National Science Foundation. http://www.communitytechnology.org/nsf%5Fci%5Freport/report.pdf.

Jenkins, H., Klopfer, E., Squire, K., and Tan, P. (2003). Entering the education arcade. *Computers in Entertainment*, 1(1):17–17.

National Science Foundation (U.S.), & Cyberinfrastructure Council (National Science Foundation). (2007). Cyberinfrastructure vision for 21st century discovery. Arlington, VA: National Science Foundation, Cyberinfrastructure Council. http://purl.access.gpo.gov/GPO/LPS80410.