

Geospatial Services and the Need for Transparency within the NSF Cyberinfrastructure

Daniel W. Goldberg - University of Southern California

Two complementary goals of the NSF Cyberinfrastructure Vision for the 21st Century are to create an architecture of computational services and a consolidated national digital data framework. Together, these two goals aim to provide researchers, organizations, and institutions with access to the computational resources and underlying data needed to investigate spatially-explicit research questions and to share their results with the greater scientific community. As these advancements are realized, scientists will begin to gain access to computational tools and data sets that were previously out of their reach due to the complexity, expertise, and time required for their development and/or the costs associated with their storage, maintenance, etc. For many scientists this will represent a shift in how research is undertaken, enabling them to perform studies that would have been impossible given the resources available to single researchers, while at the same time removing the usual control they have over the computational services and data sets enjoyed previously. In a non-cyberinfrastructure setting, a researcher usually has the ability to open their reference data files to check the accuracy of their underlying data or inspect the computational process to ensure reliability and correctness. The consumer-based approach to computational services and data provided by a cyberinfrastructure requires that this trust of ensured accuracy be placed in the provider of the data and/or services. Although the fundamental science behind many processes that operate on and/or produce geospatial data are mature, it is clear that critical challenges relating to accuracy and reliability arise in an interoperable international network of data and computational services because any error or uncertainty in these systems will propagate directly to the end consumer, potentially without their knowledge. In particular, data consumers will need to be keenly aware of the characteristics of the computational processes and data that may affect whether or not a particular data layer or set of spatial processes are suitable for use in any particular study or application. Likewise, data aggregators will need to be aware of similar issues before data is certified as fit-for-use and offered up to the research community at large.

It is a fact that nearly every computational process used in the spatial domain will have some error and/or uncertainty in the results it produces. Geospatial processes are often, by their very nature, highly complex systems relying on specific assumptions, data models, and reference data sets and as such are fraught with the potential for error at any stage of the process. Because of this, it is common that any specific implementation of a particular computational process developed by one group will produce output that varies from other services purporting to perform the same operation. As is currently the case with the process of geocoding, which translates postal addresses into geographic coordinates, the various services that one could use will routinely return results that differ from one another for the same input datum as well as from itself when queried at a different point in time, often quite dramatically. This is especially true for postal addresses that are difficult to process due to one or more of the common problems encountered in geocoding attempts: (1) input-error – problems with the actual text of the input

address; (2) reference data error – reference data files for an area are unavailable, out-of-date, incomplete, or inaccurate; (3) boundary issues – the true location is located on the boundary of two administrative areas or some other spatially-varying characteristic; and (4) implementation errors – programming mistakes that are either systematic or only appear in specific instances. These difficult cases often have the highest potential to reveal information about the phenomenon in question, so it is unfortunate that variability in output across the commercially available geocoding implementations increases with their prevalence.

So what, if anything, is a researcher investigating the link between a particular airborne substance with a decay distance of less than 50 m and the occurrence of a particular health outcome or activity pattern supposed to do when they query four different geocoding services with the same address and are returned four different locations varying by up to 1 km, all of which are purported to be of address-level accuracy? How are they supposed to rectify the disagreement between these supposedly high-accuracy data sources to determine a result with sufficient confidence which they can use in their study to classify a person as un/exposed to the toxin, and what ramifications will this conflicting information have on the validity of their conclusions? In the ideal case, one would investigate the metadata associated with each geocode to determine if they all used the same street segment from the same vintage of the same reference file to check for spatial interpolation errors or the selection of an incorrect segment. Then, given the likelihood that all services use different reference data, one would have to compare what they know about the validity, accuracy, completeness of the particular street segment chosen in each. This process of accuracy verification and evaluation would need to continue through all aspects of the geocode derivation in order for the researcher to determine the most appropriate output location based on what they know of how each geocode was produced. Sadly, this is typically impossible given the for-profit model of the geocoding services used in most research studies which hold reference data files and geocoding algorithms as trade secrets.

This situation illuminates the fundamental problem that must be addressed. As progress is made toward the development of an international set of geospatial tools and data sets, the developers and providers of the fundamental geospatial operations at the heart of this cyberinfrastructure will need to ensure transparency of their services such that international standards can be applied to quantitatively rate the services and/or data that are to be used and re-used in scientific studies. To enable this, the algorithms employed and reference data sources consulted must be available for inspection and documented with useful accuracy metrics, which by definition requires the creation of an open source set of freely available computational services that anyone could consult and/or compare. With such open systems and robust metrics in place, the GIScience component of NSF Cyberinfrastructure will be capable of reliably providing scientists the underlying data and tools they need to pose and investigate spatially-explicit research questions at the local, national, and international scales. Without them, researchers will be forced to continue to utilize services and/or data of varying and sometime dubious quality in their research endeavors.