How Deep to Go, How Soon, in Data and Code Sharing?

Alyssa Goodman, Harvard University

November 2009

As an astronomer, I am part of a community that wants to share its data. Our field is relatively small, and our data are typically without any economic value. Competition amongst individuals is only fierce in v*ery* a small number of "hot" or "new" areas, and the reality is that most astronomers are more interested in easy access to each other's data than they are in personal glory. (And it is not true that good astronomers only work on trendy hot topics.)

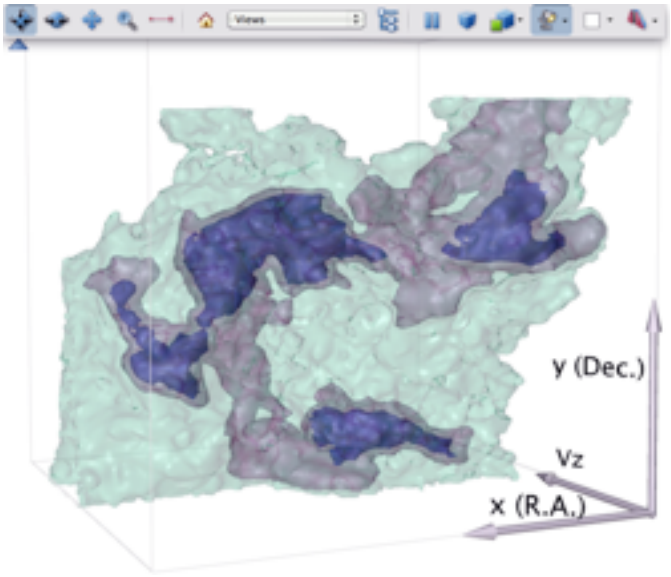
The "Virtual Observatory" in astronomy is an effort to make all astronomical data available online. Sometimes, people mistakenly think of the "VO" as a data repository, but it is not. It is an international effort aimed at creating standards and practices that allows the internet itself to function as an (albeit disorganized) repository.

In the "Seamless Astronomy" effort, I collaborate with Virtual Observatory colleagues from academia and government as well as with colleagues in industry--particularly at Microsoft Research. The goal of our project is to make the connections between data sets and the literature where data are presented, described, and analyzed as "seamless" as possible. Our goal is *not* to turn papers into perfect data archives: we do not seek to make every submitted paper 100% reproducible, right now, because that level of depth is presently unrealistic.

Instead, our initial "seamlessness" goals are grounded in what is possible according to our observations of what real, working astronomers will use. The astronomy literature is very well organized and centrally accessible online through the "ADS" literature services (http://adsabs.harvard.edu/abstract_service.html). Nearly 100% of research astronomers use ADS every day in their work. My Seamless colleagues, including the ADS team, and I believe that the data can function as a filter for literature discovery, and vice-versa, so we are taking steps to making the two realms (data and literature) more "interoperable."

There are formal links to data sets within papers available on ADS in only about 10% of cases, and the situation is not improving rapidly. We can *hope* that the prevalence of "full-depth" formal links to data will rise at some point in the future due to peer or governmental funding pressures. But, instead of just waiting (perhaps indefinitely?) for these hopes to be realized, we are working on exploiting opportunities to link relevant data to information presented in a published paper, even if it's not the exactly the explicit or raw data used in the paper. For example, we now have the technology to automatically digitize published images and re-assign metadata (coordinates on the sky) to those images using clever "astrometry" (measurement) algorithms (see http://astrometry.net/). This means we can re-access millions of images once thought to be locked up forever in the printed literature automatically. In other words, we can create metadata after-the-fact in some cases, thanks to new technology.

New tools also allow scientists to provide their data to colleagues interactively *within* a published paper. In January of 2009, my colleagues and I published the first "3D PDF" in *Nature* (see information and links at http://www.cfa.harvard.edu/~agoodman/newweb/3dpdfNews.html). Our paper's figures allow readers using standard free software (Adobe Acrobat/Reader) to and interact with data in real time (a non-interactive figure snapshot is shown here, and a sample non-copyrighted interactive file is at http://iic.harvard.edu/sites/all/files/interactive.pdf). Readers can rotate the 3D objects displayed, measure their properties, and explore a series of pre-set alternative views of the data not viewable in the printed version of the paper. *Importantly, though, the paper does not contain the "raw" data used to create the interactive figure.* There are many, many, steps involved in going from the many gigabytes of raw data astronomers take off telescopes to the highly-digested displays like the one in the "3D PDF" figure.



Typically, computer scientists represent the work of scientists as a "workflow," but the idea of a linear "flow" or even a pre-determined set of iterative steps, is not really applicable to how data sets are massaged to produce results like those shown in the "3D PDF" *Nature* paper. Instead, the "many, many steps" alluded to above can change from instance to instance and from researcher to researcher. Sure, it would be great to go to 100% provenance and be able to provide fully "reproducible" research results to colleagues. But, the reality is that even today's best, most technologically-capable, astronomers can only *describe* what software and settings were used to derive particular numbers or images. The software they use is not typically capable of automatically capturing the "settings" used in the semi-random-walk that gets researchers to a final result. And, today's astronomers simply will not take the time to manually record the dozens of decisions they have to make about seemingly-trivial settings in every program they use.

In today's environment, it is not clear that "full depth" provenance either in providing raw data or in providing "all" code and related settings is either possible or desirable. Not many researchers really want to reproduce every step of someone else's process. Sure, some future researcher *might* want to, so it's a laudable goal to make that possible at some future data, by designing systems where capturing and then extracting data/software provenance is easy. For now, though, computer scientists, software developers, and policy makers should keep in mind that *"much-improved literature-data connectivity"* is not synonymous with *"perfectly reproducible*" research.