Data and Code sharing in computational science: Two thoughts

-Ramesh Subramanian

Topic 1: Categorization and organization

Several "Science Commons" efforts are underway currently. These include open source bioinformatics, open source geo-spatial research and open neurological research data, to name just a few. One of the continuing problems affecting the utility of the "open code and data" movement is the lack of common standards for compiling and organizing code and data. I feel that there are two possible ways to address this issue. One is a top-down approach, where an effort can be undertaken to create meta-data libraries for various fields, and follow that up with a "design patterns" (Gamma, 1995) approach to create patterns that describe categorizations and solutions to various problems in the respective fields. The creation of the above meta-data or design patterns would spur development of several open-source initiatives. Different categories of code could then be developed which could be served in a "software as a service" model or cloud computing model.

A second is a bottom-up, peer-to-peer approach. In this approach no attempt is made to categorize or organize the various code or data. Instead, a client software installed in a user's computer searches other nearby peers for specific data or code. The search is propagated exponentially across the Internet until a computer that has the data or code pattern is located. This is then downloaded by the original searcher. This type of search typically functions through the use of key-word searches on documents, code documentation and data file names, and does not depend upon format or other restrictive factors. Once the appropriate files are downloaded, the researcher can then identify appropriate means of data conversion or code modification as required. This kind of an approach was used to locate, identify and retrieve research documents, presentations, code snippets and even digitized conversations in IBM Corporation for use of its 300,000 employees (Subramanian & Goodman, 2004).

In the above peer-to-peer approach, it is important to ensure the integrity as well as availability of data and code. This can be done only by maintaining an additional data store which could either resemble a Wikipedia, Google or Cyc (pioneered by Douglas Lenat) (Cycorp web site, 2009) approach. Cyc is an AI project that attempts to create a comprehensive ontology and knowledge base for common-sense knowledge. This approach could be used to organize and then find connected information (i.e. data or code), as well as map the gaps in the available information – a useful functionality. Of the three, the wikipedia approach seems most promising for two reasons: its openness and ability to represent different types of data and code, as well as the crowd-sourced security and error correction features.

Topic 2: Openness versus cultural appropriation

Making open the code and data in all research endeavors is indeed a laudable notion. However, what is so obviously "correct" to many can also have some unintentional consequences that must be examined. It has been alleged that the bio giant Monsanto appropriated data pertaining to various strains of wheat in India, and managed to acquire patents on a strain of wheat with "unique low-elasticity, low gluten properties." The problem is that wheat strains with precisely these properties have been traditionally cultivated in India for centuries, and have been documented adequately by the British Government in India in the early twentieth century. Interestingly, as noted by Vandana Shiva (Shiva, 2004), the data collection localities listed by W. Koelz of the USDA as proof of the uniqueness of the Monsanto wheat are clearly in error. For example, one of the data collection locations was listed as Marcha, at an altitude of about 3000 meters, at Latitude - 280 mm N and Longitude – 80mm E. Shiva notes that the

latitude and longitude corresponds to the plains near Shajahanpur and not to a location at 3000 meters! What this example shows is that biological information (data and processes) that have been part of a nation's heritage can sometimes be appropriated by commercial companies which can use them to create similar products. This approach, while possibly legal, raises questions on the appropriateness of such actions. In this case, it can be deduced that Monsanto appropriated data cataloged by the British many decades earlier.

Interestingly, in an attempt to comply with Article 27(3) of the TRIPS Amendment, the Indian Parliament enacted the “The Protection Of Plant Varieties And Farmers' Rights Act” in 2001. This Amendment permits sui generis protection for plant varieties, but also “recognize(s) and protect(s) the rights of the farmers in respect of their contribution made at any time in conserving, improving and making available plant genetic resources for the development of new plant varieties.” In 2006, a “Plant Varieties Registry” was installed as envisaged by the Act, and a period of three years was stipulated within which all existent varieties of 12 specified crops, already in the public domain would have to be registered. Effectively the portion of the public domain that is not registered within this period will be deemed to be absent (Prashant, 2007). This implies that some openly available biological data can be listed in trade agreements such as TRIPS specifically to restrict and narrow-down (or pigeon-hole) the scope of the data, thus affording patent protections to minor variants developed by multinational corporations. This shows that while openness of data and processes is in general useful for the development of future innovations, it could also have negative ramifications, especially with respect to inherited cultural knowledge.

Another issue pertains to research data obtained from unique and unusual biological specimens, especially isolated tribal communities in certain remote locations around the globe (such as the Andaman and Nicobar islands of India, remote islands in Indonesia or the Amazon jungles). These cultures are interesting research subjects to researchers studying the evolutions of the human genome as well as the communities' adaptation and immunity to certain diseases, etc. Opening access to such data so that more researchers can have access would have serious privacy as well as human rights implications.

Given these examples, I submit that in proposing ways to open data and code, researchers should also examine the consequences of opening up certain types of data. It would be appropriate to study the ethical and legal implications on opening data and processes (code) that pertain to certain human subjects, especially since the human subjects do not often enjoy the same privacy and human rights protections that are enjoyed by the researchers.

References

Cycorp web site. (2009). OpenCyc brings meaning to the Web. Retrieved November 16, 2009, from http://www.cyc.com/

Gamma, E. (1995). *Design patterns: elements of reusable object-oriented software.* Addison-Wesley, 1995.

Prashant. (2007, August 14). iCommons.org - Cultural Heritage and the Commons - Some vignettes. Retrieved November 15, 2009, from http://icommons.org/articles/cultural-heritage-and-the-commons-some-vignettes

Shiva, V. (2004, April 24). Wheat Biopiracy. Retrieved November 16, 2009, from http://www.zmag.org/zspace/commentaries/1921

Subramanian, R., & Goodman, B. D. (2004). Peer-to-peer Corporate Resource Sharing and Distribution with Mesh. In *E-collaborations and virtual organizations* (pp. 98-119). Hershey, PA: IRM Press, 2004.