Deluge of scientific data needs to be curated for long-term use

February 24 2010

Carole L. Palmer, a professor of library and information science at Illinois, says that data curation -- the active and ongoing management of data through their lifecycle of interest to science -- is an important part of supporting and advancing scientific research. Credit: L. Brian Stauffer

With the world awash in information, curating all the scientifically relevant bits and bytes is an important task, especially given digital data's increasing importance as the raw materials for new scientific discoveries, an expert in information science at the University of Illinois says.

Carole L. Palmer, a professor of library and information science, says that data curation - the active and ongoing management of data through their lifecycle of interest to science - is now understood to be an important part of supporting and advancing research.

"There's a lot of recognition now of the value of data as assets to institutions and to the scientific enterprise, more generally," Palmer said. "Saving only the publications that report the results of research simply isn't enough anymore. Researchers also need access to data that can be integrated and re-used in new ways. This is especially important in data-intensive science, where the power of discovery lies in applying computational approaches to large, aggregated data sets."

Palmer, who also is the director of the Center for Informatics Research in Science and Scholarship at Illinois, said that researchers need to start thinking about data-management requirements from the very beginning of their projects, and to think in terms of a data set's lifecycle.

"Data curation emphasizes the lifecycle - managing and preserving data for the long term, and that process begins long before data are generated," Palmer said. "Data curation needs to be introduced at the proposal stage to make sure a viable data-management plan is in place at the outset of a project."

The biggest difficulties in collecting, curating and managing large amounts of data over the long term have to do with cost and labor.

"Most organizations have serious problems with data management because it's expensive to do systematic curation, which includes documenting the context in which data were generated or derived, including the instruments involved, the protocols and such," Palmer said. "But that also requires caring for the data and making them available to other scientists. It takes serious commitment and investment."

One rationale in favor of collecting and curating data is the issue of replication and validation of a research project's conclusions, which is very important to scientists.

"It's about doing good science, and ensuring the science is reproducible, but it's also about finding ways to bring data together from different sources and using them in new ways," Palmer said. "To replicate a study or re-use data you have to know where a data set came from and how it's been processed. Tracking all the context and transformations is part of the curation process."

But even with the ubiquity of the Internet and search engines, data stored online have proven to be much more ephemeral than data preserved in print.

"Digital content, including digital data, is much more vulnerable than the print or analog formats we had before," Palmer said.

To those who would say publish it on a Web page and let Google cache the page for posterity, Palmer argues that businesses don't have the orientation necessary for curating and preserving information for the really long term - say, for hundreds of years.

Research libraries, on the other hand, have this mission and always have been committed to this.

"The common perception is that keeping information online keeps it alive," Palmer said. "But someone, somewhere, has to maintain it and make it accessible and usable for researchers. It's not wise to rely on publishers or other commercial entities that have never really been in the business of preservation. Businesses can go out of business, and they're driven by commercial interests."

"So just assuming that if it's online, it's accessible - well, it's accessible until it's not."

Palmer says there's also a lot of work involved in selecting, appraising and organizing data to make them accessible and interpretable. Just because a page can be saved in a Google archive doesn't mean it's going to reveal itself to an inquiring researcher.

"Google is great for the front-end, for getting a thread into something," Palmer said. "But it doesn't offer you the context that's important for aggregating and presenting data for research purposes. Bringing data sets together, organizing them and linking them to related documents and tools so that scientists can work with them is something very different than search engines scouring the Internet for something related to a search term."

Palmer says there's also a public-access argument to consider, especially if the research was funded using public dollars.

"There's a growing interest in open data, making data part of the outcomes of research that are disseminated to the public rather than keeping them as private possessions of the researcher," she said.

This trend also can expand who's involved in the scientific research process. For example, amateurs have discovered new stars using public archives of astronomy data, and citizen scientists are better able to contribute observations and measurements on climate, plant growth and other areas beneficial to longitudinal environmental studies.

"There's a lot of new activity in data curation, but the field is still in its infancy," Palmer said. "We're just beginning to do the research needed to guide how we build large-scale, multidisciplinary data repositories and collect and manage data in ways that add value and promote sharing and integration across laboratories, institutions and disciplines."

In terms of preparing a workforce, "We've also begun to train data curators and to build up professional capacity," Palmer said.

"The bottom line is that many very talented scientists are spending a lot of time and effort managing data. Our aim is to get scientists back to doing science, where their expertise can make a real difference to society."

The Center for Informatics Research in Science and Scholarship at Illinois will receive about $2.9 million as a partner on the Data Conservancy project, a $20 million initiative led by Sayeed Choudhury at the Johns Hopkins University Sheridan Libraries. The five-year award, one of the first two in the National Science Foundation's DataNet program, will fund developing infrastructure for the management of the ever-increasing amounts of digital research data.

The Illinois team is conducting studies of scientists' data practices and needs, and analyzing how to best represent complex units of data in the repository, while also further developing their professional training programs in data curation.

Provided by University of Illinois at Urbana-Champaign

Citation: Deluge of scientific data needs to be curated for long-term use (2010, February 24) retrieved 25 April 2024 from
https://phys.org/news/2010-02-deluge-scientific-curated-long-term.html