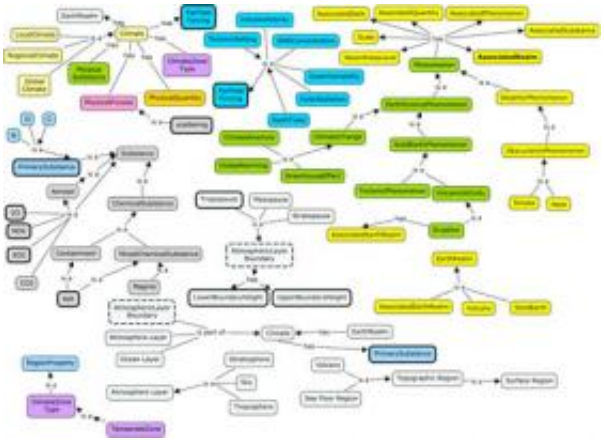
The eScience revolution: Creating semantic Web platforms for massive scientific collaboration

October 1 2009



The Semantic Web technology being creating in the Tetherless World Research Constellation will allow scientists, educators, and people around the world to access data on a variety of topics all in one place, bringing together scientific data in unprecedented ways. Represented here are just some of the areas that would intersect during a search for data on the Earth's atmosphere. Credit: Rensselaer/Peter Fox and Deborah McGuinness.

(PhysOrg.com) -- Web scientists at Rensselaer Polytechnic Institute will use the World Wide Web to compile and share scientific data on an unprecedented scale. Their goal is to hasten scientific discovery and innovation by enabling rapid and easy collaboration between scientists, educators, students, policy makers, and even "citizen scientists" around the world via the Web.

Funded by $1.1 million in American Recovery and Reinvestment Act funding from the National Science Foundation (NSF), the research seeks to break science out of the hallowed halls of the laboratory and place it in the hands of the people.

"We want to provide a toolkit for scientists and educators that allows them to gain access to data from a variety of sources and, importantly, outside of their direct area of expertise," said Peter Fox, the principal investigator for the project and Senior Constellation Professor in the Tetherless World Constellation at Rensselaer. "Right now there are many scientists, educators, and policy makers who want to use other's scientific data, but they don't know how to find it, how is was collected, and even how to read it." Fox notes that with the increased specialization of most scientific research, even people in closely-related fields currently struggle to interpret the data of their contemporaries. These scientific language barriers, he said, can hinder the pace of new discoveries.

The new toolkit will have a foundation in Semantic Web technology. On the Web, semantic computer code (known as ontologies) provides underlying meaning and links to the information that is presented on a Web page to your computer, smart phone, or other Web-enabled device. Current technology involves flat words on the screen, for example "climate change," that require a human to interpret the words and then manually move on to another Web site for additional information. Web technologies based on semantics, however, would enable the computer to provide its own underlying meaning to the words, and provide links to related Web sites, nonprofit organizations, upcoming Senate bills, or even related photos stored on your computer. In the case of semantic data, the computer can configure, coalesce, and interpret data from millions of different sources instantly without the need for human intervention.

"Semantic technologies lower the barrier of entry to do science," said co-principal investigator on the project and Senior Constellation Professor Deborah McGuinness. "With semantics, we can bridge the gap between the question that someone wants to ask in their limited scientific vocabulary and the extreme complexity of the underlying data." An individual's vocabulary and scientific understanding will no longer have to correspond to the level of their scientific discovery, according to Fox and McGuinness.

Fox, McGuinness, and their counterpart on the project, Senior Constellation Professor James Hendler, will use semantic ontologies to build customizable Web sites. Each Web site will be familiar, understandable, and navigable to its end user depending on the level and type of expertise. Behind the simple façade of the Web site will rest billions of pages of data all semantically tagged and ready to be accessed and interpreted by the computer. The user needs only to type a question, and it will be answered using data input by other users around world. The researchers also plan to create plug-in applications for commonly used data software such as Excel that adds access to the data in a format that is familiar to the end user.

All of their semantic coding will be open source, making it available to others on the Web seeking new ways to share data.

"We want to accelerate the growth of community knowledge," McGuinness said. "We want to encourage others to look at the data, interpret the data in their own ways, reuse the data, and even verify the data."

Fox, McGuinness, and Hendler see the technology helping to lead a revolution in the citation and, possibly, review of scientific data. Much like Wikipedia, the data on their Web sites and technologies will be viewed and used by users from leading scientific experts to elementary

school teachers and all those reviewers will be able to comment and cite the data.

"There will be extensive new opportunities to review the data," Fox said. "It may not be a traditional peer review as is the custom in scientific publication because many people will not be experts, but each user will bring a very legitimate point of view to the data, particularly when they use it in new and different ways." Thus, a school teacher could make a discovery on sea level change that an oceanographer may never have found.

The ease of access to the data will also allow other scientists to quickly reproduce and verify a data set. Often in a scientific paper, there will be a scientific figure or image that represents a data set. Raw data is rarely presented, making it extremely difficult for another scientist to pick up where another left off or even reproduce the results, according to Fox. The new semantic technology will mediate access to the raw data and in a vocabulary that the end user can understand.

In addition to ease sharing data, the semantic technologies will also allow for ease of citation when using data created by someone else. Access to certain data sets can be controlled and with semantic tags attached to the data of their source, and users can easily give credit to the original creator of the data that they are utilizing, while data creators can track exactly who is looking at their data. "For the first time, we could see scientists citing online services in peer review journals," McGuinness said.

Semantic e-science is an area of unique specialization within the Tetherless World Research Constellation, which is comprised of "star" faculty who mentor up-and-coming faculty, graduate and undergraduate students in fields ranging from computer science to informatics. Their collective research and teaching efforts center on the emerging field of

Web Science and seeks new ways to understand and harness the inner workings of one of the most powerful research, social, and commerce technologies of our time.

Source: Rensselaer Polytechnic Institute (news : web)