CI-Server: Towards a Collective Scientific Knowledge Environment

Aída Gándara
Department of Computer Science
The University of Texas at El Paso
500 W. University Ave
El Paso, TX 79968
agandara1@miners.utep.edu

Paulo Pinheiro da Silva
Department of Computer Science
The University of Texas at El Paso
500 W. University Ave
El Paso, TX 79968
paulo@utep.edu

ABSTRACT
A well known challenge in scientific computing is the flow of information sharing in support of scientific research. The issue is further exacerbated when research is done collaboratively because more people need access to the same information, often simultaneously. The Web has emerged as a popular solution for enabling discovery and sharing of information between people and applications. Unfortunately, use of web-based technologies is not always easy. For example, two sites can be so different that searching for information on one site may have little similarity to the other. In this paper, we describe an environment that is focused on facilitating the sharing of information for scientific research. Through a server that we call the CI-Server, we support the collection of structured and unstructured scientific data as well as discussions about the data. A client-based API (CI-Client API) was created to enable access to the CI-Server from within scientific applications. Thus, scientific collaboration is supported by allowing the scientist to work with their tools while sharing over the Web instead of the traditional method of having the scientist learn new environments. As a result, scientists are not forced to learn specific web server or portal environments because the utilities that they are accustomed to using have the 'know-how' to access the knowledge collected at the CI-Server. CI-Server has been used in support of scientific activities in the areas of environmental and geo-sciences as part of a NSF-funded Center for Cyber-Infrastructure, CyberShare.

Author Keywords
Semantic Web, Scientific Collaboration, Collective Knowledge, Collective Intelligence System

ACM Classification Keywords
K4.3 Organizational Impacts: Computer-supported collaborative work

INTRODUCTION
Sharing information through web pages, wikis and social networking tools shows how useful the Web can be in promoting collaboration. Use of the Web, though, has introduced some issues that must be addressed when working in a scientific environment. For example, the ability to share information must be seamless within the context of a scientist's environment. Most scientists we have worked with require an intermediate step where they must either upload artifacts within a portal or request uploads from a webmaster.[1] The need for waiting to publish or publishing within the framework of a web page or portal can hinder communication between scientists. Privacy is another issue. Scientific information is sometimes sensitive and scientific teams might have concerns about its accessibility over the Web. Finally, as discussions occur about scientific research, there is a lot of user generated information in the form of emails, chats, blogs, etc. that is lost or difficult to relate to research related artifacts. Making use of the power behind Web technologies to collect user-contributed content and machine-gathered data enables the creation of emergent knowledge that can be used to answer more research related questions. [2]

In this paper we introduce an environment to help support scientific teams in collecting research-related information and discussions over the Web. The CI-Server is a content management system that has been designed to collect knowledge about scientific research. The CI-Client is a client API that has been created to embed knowledge within the tools of the scientist to support sharing of information and enable discussions about scientific data. The goal is to enable scientists to work within the context of the tools they are accustomed to using to share, work on and discuss scientific research without having them learn the context of portals and websites. Thus, any client tool that embeds the CI-Client technology is interoperable with any server that embeds the CI-Server technology.

MOTIVATION
Scientists working collaboratively often have a challenging task when it comes to sharing information. When data is produced there is often a challenge of making research related information available to fellow collaborators. For

example, once a geologist derives tomographic images of the Earth, he or she must be concerned with its availability to fellow scientists. For instance, the scientist may need feedback about the quality of the image from other members of his or her team. Furthermore, when discussions occur about the data or the process for collecting the data, the team is left with a collection of emails, chats, etc. containing related discussions and very often critical meta-information about the data with no techniques to support the collection, organization and understanding of these discussions.

One solution to sharing information has been to publish content over the Web. In terms of accessibility, the Web provides data to over 6 billion users [3]. By publishing information over the Web, more scientists have access to web technologies for solving problems as a team. There is a bottleneck in sharing information when scientists must wait for permission or expertise, as is the case with many web sites; or must upload the content in a portal that may not align with scientific related applications a scientist is accustomed to using. For example, the scientist may use a system that utilizes a map to view and select scientific data whereas the portal may require the user to search in a tree structure interface.

Collaboration is further supported by tools like email, chats and, more recently, social networking tools. Numerous tools exist to support socializing over the Web, e.g. Facebook[1], Second Life[2], MySpace[3]. In these environments users can hold discussions on any subject and share additional information like photos, videos and personal or business related attributes. In relation to scientific research, these social networking tools do not provide any support to relate the data created during scientific research to the information generated in discussions. For example, there is no requirement in Facebook for a scientist to add specific scientific information related to a discussion. As a result, if there is a need to reference the discussion or provide discussion information to justify research ideas, there is no reliability to finding relevant conversations.

A Collective Knowledge System,is a web-based software system that functions to collect knowledge as well as build upon that knowledge through reasoning over the collected information.[2] An important component to implementing such a system is the ability to structure information and link it to unstructured information in order to relate meaningful attributes to data. For example, the collection of conversations can be collected as unstructured segments of text. In the context of scientific research, these are actually annotations to scientific data, inputs, artifacts or computational devices. Thus, by applying structure like labels or tags to this unstructured text, the knowledge system now has additional information related to scientific research. Furthermore, the Collective Knowledge System now has the ability to search and make inferences based on the text that has been related to the research. In this way, the collection of information can be used to infer new information that was not available before, thereby creating a collective intelligence.

Providing support within scientific applications to help build the collected knowledge enables systems, and therefore scientists, to benefit from the collective intelligence within the context of the tools they use daily, not using disjoint web servers or portals.

THE CI-SERVER APPROACH

CI-Server is a web application that aims to collect knowledge about scientific research including structured and unstructured artifacts by reaching out to tools and environments often used by scientists. These artifacts are built by client applications and published on the server. Capabilities of the server are built upon the collected knowledge to create additional inferred information. The CI-Server manages the following:

Privacy – groups and users that have access to resources collected on the server. Access to resources is controlled per user and per group. Furthermore, the information can be made public over the Web, or accessed privately, through services.

Resources – collective knowledge, both structured and unstructured, that is created by scientific applications and published on the server. Structured knowledge may include ontologies, abstract workflows, and provenance[1]. Unstructured knowledge may include scientific data, word-processor documents, spreadsheets, among others.

Collaboration – content created by discussions within the framework of scientific tools. This information is stored as structured data related to resources. The Server defines the link between scientific and social information.

Inferred Knowledge – new information that is created as a result of knowledge published on the CI-Server. For example, as scientific discussions are published on the server, the server can keep track of agreements or disagreements, inferring a percentage of either.

THE CI-CLIENT COMPONENT

The CI-Client is an API that can be integrated into scientific applications to provide access to CI-Servers. The tool enables applications to seamlessly access and create knowledge on the CI-Server within the context that the scientist is accustomed to. The CI-Client provides the following functionality:

Authentication – conforms to the needs of the group accessing the data on the server. If the group must access information using user authentication, then the group can

[1] http://facebook.com
[2] http://secondlife.com
[3] http://myspace.com

rely on a needed level of privacy. Similarly, all or partial data can be made public for access to all Web users.

Access to Resources – provides access to the resources available on the CI-Server. In particular those that are created by the client application or might be related in scientific research.

Access to Collective Intelligence - Initially, the main challenge of the CI-Client is to provide access to the information that is generated by the client application and published on the server; but access is not limited to that. The benefit of having access to the server is access to the collective intelligence. As data and discussions are related to scientific research, clients should be able to build information and access additional knowledge that resides at the server, including inferred knowledge.

CI-Server interoperability - The generality of the CI-Client must provide access to multiple CI-Servers and be able to relate and integrate domains found on distinct CI-Servers.

A CI-SERVER FOR WORKFLOWS

An initial implementation of the CI-Server has been developed in Drupal[4]. Drupal was chosen in order to take advantage of basic levels of security, web-service support to access server content, extensibility and messaging. The system is currently in use by a scientific team focused on adding provenance capture to their scientific research. The initial task for this team is to understand the process used to collect scientific data. The team is working from a pre-existing workflow representation of a known algorithm and must work to modify the workflow to reflect added parallelism. The difficulty, in part, arises from the fact that the experts that captured provenance for the original algorithm are not the same experts that migrated the process to a parallel architecture. Thus, the team must first agree on the steps in the process then agree on the provenance capture. All access to the server is via http or xml/rpc-based web services enabled in the workflow editor through the CI-Client which has been developed in Java.

The workflows, which are OWL[5] ontologies, were created by the workflow editor and published on the CI-Server. To test the flexibility of this environment, the workflows themselves are accessible over the Web as URLs, but the discussions are kept private to the scientific team. Therefore, the users of the CI-Server require authentication to access discussions and were each assigned an individual username and password.

Discussions are currently supported with messages sent to the server. These are collected at the server and related to the workflows. Scientists, from within the workflow editor, are able to request the messages that have been collected for a workflow, as well as add their own comments about the workflow for other team members to see.

The CI-Client manages server connections and specific CI-Server information. In this way, the interoperability between the CI-Client and CI-Server is transparent to the scientific application. Furthermore, the CI-Client supports access to multiple CI-Servers enabling integration of information.

Current efforts are focusing on 1) the building of collective intelligence at the CI-Server and 2) bulk uploads of scientific information occurring seamlessly from the scientific application.

RELATED WORK

Publishing scientific information is not a new process. Several environments are available, most in the form of portals, where scientists are able to upload and download scientific information in the form of structured or unstructured data. For example, the Earth Scope portal [9] provides access to an infrastructure of instrumentation devices, data, observatories and other Earth related information. As a data portal, EarthScope provides a Google-Maps based interface to search for and download scientific data, e.g. seismic data, GPS data. As a result, users of the portal must follow EarthScope conventions to download useful scientific information. A team of scientists might find good use for such data but they are limited in actually using such a portal to share their own information since the EarthScope portal is not open to user publishing. Another example is the BioPortal, a portal providing web-based access to a library of biomedical ontologies.[4] BioPortal is open to user contributions of ontologies. Furthermore, this portal supports social-based user contributions in the form of feedback and user evaluations. Since all ontologies are publicly accessible, there is limited team privacy. One interesting feature provided by the BioPortal is versioning. Although the current implementation of CI-Server does not support versioning, we are aware of its importance for supporting individual suggestions and subsequent changes to the workflows. Nevertheless, publishing information on these sites requires understanding how the portal works, where menus are located and publishing information specific to the rules of the portal. The more portals a team of scientists might need, the more work needed to understand each portal.

myExperiment [5] is also a portal for discussing scientific workflows; similar to the initial implementation of the CI-Server. As part of an initial evaluation of myExperiment, De Roure, et. al. discuss the requisite features necessary to support the 'Social Virtual Research Environment' that myExperiment is based on. A similar evaluation of CI-Server aligned with myExperiment will serve to compare and contrast these two projects. The first characteristic necessary is to facilitate the management and sharing of

[4] http://drupal.org
[5] www.w3.org/TR/owl-guide

Research Objects. Both tools focus on the management of workflows primarily but allow for access or links to other types of objects. Although workflow support is the initial implementation of the CI-Server, the research objects are open to any type of structured or unstructured data. One difference here seems to be in the focus of supporting additional research objects. myExperiment is building functionality to support client-like knowledge at the server level by building into myExperiment knowledge for visualization or workflow execution. CI-Server is focused more on allowing scientific applications to work with the data at a client-based level and building collective intelligence at the server, based on the information collected. We are aware of the fact that some client information may be necessary to enable added knowledge but we are expecting to rely on structured information to enable such functionality. Second, is support for a social model. By far, the strength of myExperiment in this quality exceeds that of the current implementation of the CI-Server. The plans are to support more elicitation of knowledge through the CI-Client. The third characteristic is to provide an open-extensible environment. myExperiment does this with an API and web-services. One of the main features of the initial implementation of the CI-Server was to use Drupal. Drupal is extensible and open. First, the CI-Server is a module written in PHP that is enabled through the Drupal application. To implement a CI-Server, the module can be installed and extended to support additional features. Furthermore, through the CI-Client, client tools can build knowledge into any scientific or even social environment. We have plans to target common social networking tools as a proof of concept. Our current work on the server is focused on building added knowledge to the system through reasoning based on collected knowledge. As users add more information, the CI-Server is expected to extend itself. Finally, with the ability of the CI-Client to access and integrate with multiple servers, client tools can leverage inter-disciplinary knowledge available at different CI-Servers. The final feature, platform to action research, is supported by myExperiment through the features of the tools built into the portal, e.g. the executable workflow tool. Users can leverage the executable workflow functionality to execute and experiment with the tool from within myExperiment. CI-Server's action research is based on inference capabilities that occur as a result of the data that is generated from the collected data. One interesting characteristic of the myExperiment system is a recent effort to model the system using an OWL DL ontology[6]. Structured information like ontologies, we believe, will be instrumental in the success of the CI-Server's abilities to collect and infer knowledge about the system. Thus, we believe that using similar models to understand clients and related characteristics like collaboration will be important to the CI-Server efforts.

CONCLUSION

Tools that support collaboration for scientists must support a flow of information sharing that enables scientists to work together. Tools that require scientists to publish on websites or portals cause bottlenecks in collaboration because scientists must adapt to those environments which may not be what they are accustomed to or may involve learning many different environments. The CI-Server in combination with the CI-Client establish a framework where applications can seamlessly connect to servers to share and use published data via the tools that scientists are use to using to conduct scientific research.

As scientists share more knowledge about their research on the server, for example by participating in discussions or publishing scientific data and related artifacts, the CI-Server can learn about the research and build knowledge to help answer more questions. In this way, the CI-Server framework is becoming a collective scientific knowledge environment.

ACKNOWLEDGMENTS

This research was partially funded by NSF Grant #HRD-0734825. Any opinions, findings and conclusions or recommendations expressed in the paper are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

1. Pinheiro da Silva, Paulo, Salayandia, Leonardo, Gandara, Aida, Gates, Ann Q. CI-Miner: Semantically Enhancing Scientic Processes. To appear in Earth Science Informatics, Springer. 2009. DOI 10.1007/s12145-009-0038-3.
2. Gruber, T. 2008. Collective knowledge systems: Where the Social Web meets the Semantic Web. *Web Semant.* 6, 1 (Feb. 2008), 4-13.
3. Usage and Population Statistics. Internet World Stats. 2009. http://www.internetworldstats.com/stats.htm
4. Noy, Natalya F., Dorf, Michael, Griffith, Nicholas, Nyulas, Csongor, Musen, Mark A. Harnessing the Power of the Community in a Library of Biomedical Ontologies. Workshop on Semantic Web Applications In Scientific Discourse (October 2009).
5. De Roure, D., Goble, C., Stevens, R.: The design and realization of the myExperiment virtual research environment for social sharing of workflows. Future Generation Computer Systems 25 (February 2009) 561-567.
6. Newman, D.R., Bechhofer, S., De Roure, D., myExperiment: An ontology for e-Research. Workshop on Semantic Web Applications In Scientific Discourse (October 2009).