SRG: A Digital Document-Enhanced Service Oriented Research Grid

Geoffrey Fox[1,2], Ahmet Fatih Mustacoglu[1,2], Ahmet E. Topcu[1,2], Aurel Cami[3],
[1]*Community Grids Lab, Indiana University, Bloomington, IN, 47404, USA*
[2]*Department of Computer Science, Indiana University*
[3]*Department of Biomedical Informatics, University of Pittsburgh*
{ gcf, amustaco, atopcu}@cs.indiana.edu, camiau@cbmi.pitt.edu

Abstract

We describe a new framework for building a system that consists of tools and services for supporting Cyberinfrastructure based scientific research. This system, called the Semantic Research Grid (SRG), integrates a number of existing online research tools (social bookmarking, academic search, scientific databases, journal and conference content management systems) and aims to develop added-value community-building tools that leverage the semantic analysis of digital documents. We discuss the design, the overall architecture, and the implementation of SRG, and provide a roadmap of the future work in this project.

KEYWORDS: Cyberinfrastructure based scientific research, annotation, academic search, scientific databases.

1. Introduction

In recent years there has been a rapid development of tools and services aimed at fostering online collaboration and sharing between users and communities. Blogs (blogger.com, Google Blog), Wikis (Wikipedia, WikiWikiWeb, Wikitravel), Social Networking Tools (MySpace, LinkedIn), Social Bookmarking Tools (del.icio.us, Flickr, YouTube), Syndication Feed Aggregators (Netvibes, YourLiveWire) and other related tools are quickly being embraced by an expanding user base. The term "Web 2.0" [1] is now a widely accepted term representing this wave of new Web-based tools and the belief that they indicate a qualitative change in today's Web. This change is also apparent in the domain of scientific research, with the recent creation of a number of online tools that enable the annotation and sharing of scientific content, such as CiteULike [2], Connotea [3], and Bibsonomy [4].

These developments overlap with ongoing efforts to exploit Grid architectures based on Web services [5] for supporting international scientific and engineering research teams in sharing large data and compute resources (i.e., creating a Cyberinfrastructure for e-Science [6, 7]).

Significant advances have also taken place in the areas of digital libraries and academic search. Domain specific academic search tools, such as CiteSeer [8], or general ones, such as Google Scholar [9], have enabled open, fast and easy access to vast online repositories of linked scientific documents.

In this paper, we describe a framework that is motivated by the above concerns and aims to develop a community-centric platform of tools and services that integrate the major existing annotation tools, academic search tools, and scientific databases into the Cyberinfrastructure based scholarly research. These tools and services, collectively called the Semantic Research Grid (SRG) [10], are backed by databases which store user and community specific data and metadata and have been configured into three applications: (1) A model for scientific research which links both traditional simulations and observational analysis to the data mining of existing scientific documents; (2) A model for a journal web site supporting both readers and the editorial function; (3) A model for a natural collection of related documents such as those of a research group or those of a conference.

The rest of this paper is organized as follows: Section 2 gives an overview of the existing online tools that form the basis of SRG and explains how they are used in this system. Section 3 describes the design principles and the overall architecture of SRG, expounds the various technologies and software packages used in developing this system, and details of the SRG modules. Section 4 presents a roadmap of the future work in this project.

2. Overview of Existing Tools

2.1 Annotation Tools

Perhaps, the best known annotation (or, social bookmarking) web site is *del.icio.us* (henceforth referred to as Delicious), a tool designed to enable the annotation

and sharing of URLs. A number of other annotation tools are now in widespread use; they support annotation and sharing of a variety of resources, such as photos (Flickr), videos (YouTube), books (LibraryThing) and goals (43things). In particular, there are several online tools specializing in the annotation of scholarly publications, including Connotea, CiteULike, and Bibsonomy [11]. The core service offered by these annotation tools is the capability that allows users to quickly annotate their favorite resources (URLs, photos, or citations) using a small number of *tags* (keywords) and to share their tagged content with other users.

Tagging represents a significant shift in the *metadata creation* methodology. The new approach of metadata creation, namely *tagging*, puts the task of metadata creation in the hands of general users. Among the benefits of tagging are: (a) the ease of use and access of the tagging tools; (b) the ease of discovering new content; (c) the support for the creation of niche communities. The shortcomings include: (i) the lack of a standard set of keywords; (ii) the difficulty of dealing with misspelling errors, synonyms, and acronyms, which are commonly found in tagging; (iii) the difficulty of inferring hierarchical relationships between tags (i.e., creating a taxonomy).

Each social bookmarking tool can be described in terms of: (a) A model of *data* and *metadata* adopted by the tool; (b) A *user interface* that allows users and groups to subscribe to the service, manage their tagged content, share it with other users, and discover new content; (c) An *input/output interface* that allows the data and metadata to be exported to various formats or applications, and enables programmatic interaction with the system. In the accompanying technical report [12] we give a detailed description of the above features for Delicious, CiteULike, Connotea, and Bibsonomy.

2.2 Academic Search Tools

The advent of the World Wide Web has led to the creation of a number of digital databases of scientific content. These databases use one of two main data acquisition methods: (i) manual insertion by volunteers (e.g., DBLP); (ii) automated harvesting of open-access databases, home pages of authors, web sites of the publication venues, and so on (e.g., CiteSeer). Both methods may be complemented with user submissions.

In our framework, we focus on the major open-access academic search tools that use automated methods of acquiring and analyzing scientific documents. These tools are discussed next.

CiteSeer: CiteSeer was introduced in 1997 by Giles et al. [8]. As the first tool in this category, CiteSeer is probably also the best known, especially in the field of Computer Science, which is its specialization domain. The core feature of CiteSeer is *Automated Citation Indexing*, a method for the automated extraction, parsing and indexing of the citations contained in a paper and of the context of these citations in the paper's body.

Google Scholar: Google Scholar (GS) first became public in 2004. The methods for collecting and analyzing documents used by GS are similar to those of CiteSeer. Note that CiteSeer is both a search system and a digital library having currently more than 800,000 full-text documents in its repository, while GS is a search system which attempts to find and display the URLs that point to the full-text versions of the query results. Unlike CiteSeer, GS aspires to be a "single place to find scholarly materials" covering "all research areas, and all sources" [13].

Windows Live Academic: Windows Live Academic (WLA) is the latest addition in the area of open-access academic search tools; it became public in 2006. Its objectives are similar to those of GS, but unlike GS it has revealed the list of the covered publishers and venues. The initial version of this tool, has been shown to suffer from the same issues of *coverage* and *accuracy* discussed above for GS [14]. Another drawback of WLA is that, unlike CiteSeer and GS, it does not yet provide *citation indexing*.

We achieve integration with the above academic search tools by building wrappers around them. It has been suggested [15, 16] that, for specific user categories, the "one stop shopping" or "one size fits all" approach of GS and WLA can't be an alternative to specially crafted portals integrating data from various sources. We share this belief and envision that these tools will have two main roles in the usage scenarios of our system: (1) They will be used to *seed the creation of a community* (e.g., the papers of a research group, the papers on a chemical compound, etc.). (2) They will be used to *extract the citation count* of scientific papers. Due to their global nature, GS and WLA are uniquely positioned for providing this kind of service. We anticipate that such counts will also need to be refined by community-specific tools.

2.3 Scientific Databases

Several excellent open-access scientific databases, such as PubMed, PubChem, and Science.gov, have been created over the years. These databases constitute the "deep Web" and have been estimated to contain 400-500 times more public content than the "surface Web" [17]. Since the deep Web is largely invisible to current search engines (including academic ones), this wealth of

information has not been integrated with the online research tools. Our system intends to tap into this wealth of domain specific information by focusing initially on the field of Chemistry.

In summary, the Semantic Research Grid aims to develop a set of new tools and services that aggregate information from a variety of sources (i.e., "mash-up" tools) and provide added value to communities of researchers. Next, we provide a detailed discussion of the techniques that will be used in developing these tools.

3. Design and Implementation of the SRG

3.1 System Design and Architecture

We have followed Web 2.0 design patterns [1] in designing the SRG system. Below, we list these patterns and discuss how they were applied in designing SRG:

Delivering services, not packaged software: SRG is a collection of tools and services that can be accessed over the Web (either through a user interface or programmatically through Web services). It will evolve by introducing new features; still its users won't have to install new versions of the software.

Producing hard-to-recreate data that gets richer as more people use the system: By combining data from a variety of sources, SRG will create added-value data and metadata generated with specific communities in mind.

Harnessing collective intelligence: Through its integration with the social bookmarking tools, SRG can leverage data and metadata from a large number of researchers. Moreover, the system can handle both individual users and groups of users, and supports sharing and collaboration between group members.

Leveraging the long tail through customer self-service: The term "long tail" here refers to the concept formulated by Anderson [18] that non-hit products can collectively make up a market share that may exceed the relatively few current hits, bestsellers or blockbusters, provided the store or distribution channel is large enough (this business model is leveraged for example by Netflix or Amazon.com)[1]. SRG aims to support research communities, such as the members of a research project, a group interested in a particular chemical compound and so on, by allowing them to create system accounts and to use the community-building tools for their specific usage scenarios.

Software above the level of a single device: Currently, the SRG user interface runs in a browser. However, because of its layered design and the use of J2EE technology, system front-ends for other devices, such as PDAs, can be developed at low cost.

Figure 1 shows the overall architecture of the SRG system. This system consists of three main layers: (a) the *client* layer; (b) the *Web* layer; and (c) the *data* layer. The client layer is made up of Java Server Pages (JSP) which are translated into Servlets by an Apache Tomcat J2EE Web container and generate dynamic content for the browser. The client layer communicates with the Web layer over the HTTP protocol through SOAP messages encapsulating WSDL-formatted objects. The Web layer consists of several Web services who handle communication with the existing online tools. The Web layer communicates with the data layer through JDBC connection. Finally, the data layer is composed of several local and remote databases.

3.2 Key Design Issues

We now discuss several key issues in the design of the SRG system:

Users and Profiles: The SRG system supports individual users and groups of users. Users' personal information and the login information for bookmarking web sites are accessible through the user's profile. Users can access and modify their profile settings at any time; while logged in users can: (a) Change their system password; (b) Update their profile including the full name, email address and the username and password for the annotation web sites; (c) Make requests to subscribe to any available group.

Group Administration: There are three types of users in the system: Super Administrator (SA), Group Administrator (GA), and (regular) User. There may be more than one SAs; an existing SA can add other SAs to the system. Each group has at least one GA who is appointed by an SA. When a new group is created, the user who requested the group creation becomes GA for this group. Users can make requests to subscribe to any group. GAs confirm/deny the request(s) made by users. Users are allowed to belong to more than one group.

Access Rights: Users create Digital Objects(DOs) in several ways: (a) using annotation tools (Delicious, CiteULike, Connotea); (b) using search tools (GS, WLA); (c) manually, through "Insert New DO" interface. For each DO record, there are three types of access rights: *Read* access right, *Write* access right, and *Delete* access right. Users who have Read access for a citation can read that DO. Only users who have *Write* access for a DO can update that DO. Delete access is required for deleting DOs.

These access rights are defined with respect to three kind of users: *Owner* who is the user that initiates the DO

[1] The term "long tail" is also used in statistics to describe certain statistical distributions.

metadata creation; *Group* which is the group to which the owner belongs; *Other* users.

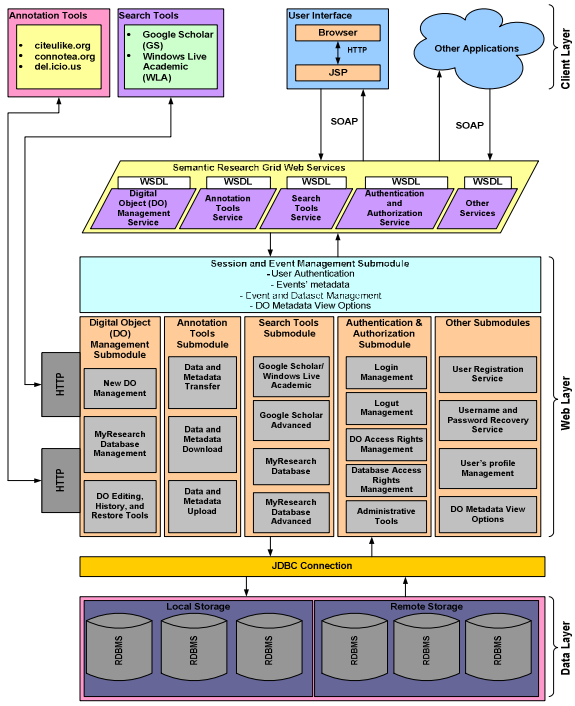

Figure 1. Semantic Research Grid Architecture

User Session: Due to the stateless nature of HTTP, a number of alternative mechanisms have been developed for applications that need to maintain a conversational state. The *HTTP session API*, which is a component of the Java Servlet specification, provides a mechanism for web-based applications to maintain a user's state information. This mechanism, which is called *session,* is usually associated with a user and supports the management of the user's state information on the server side. A session is represented by an *HttpSession* object, which stores and provides access to the user specific data.

Consistency Model: In a collaborative environment, people work together and share some resources to achieve common goals [7]. In such systems, resources are vulnerable to user mistakes. To provide consistency and to avoid undesired changes in the system, it is necessary to have a mechanism for restoring the system to any previous state. *Versioning tools* for software development, such as *Concurrent Versions System* (CVS) or *Subversion* (SVN), and *Wikis* are well-known examples of collaborative systems that provide such mechanisms.

The SRG system is a collaborative environment that allows multiple users to create, and manage a common set of citations. Data and metadata can be transferred into SRG from different online sources, such as bookmarking web sites, academic search tools, scientific databases, and journal and conference content management systems. Users are allowed to overwrite or modify existing citations; this may lead to various issues. For instance, one user can create an entry for a citation downloaded from Delicious (including tagging metadata). Later, a second user can try to insert into the system the same citation found through a Windows Live Academic search. The second user could choose to overwrite the existing citation, thus causing the tagging information for that citation to be deleted.

To allow such issues to be fixed, we have developed our consistency model based on the concepts of *event* and *dataset* [19]. An event is commonly defined as the act of changing the value of an attribute of some object [20]. Storing all the events about an object, allows users to review and undo these events. In the consistency model of the SRG system, we have adopted the view of an event as a time-stamped action on an object. We distinguish between two types of events: *major* events, and *minor* events. The insertion of a new citation record into a database and the deletion of record from a database are considered major events. Any update or modification of an existing citation record is considered a minor event. A detailed discussion of our event model can be found in [19].

Every event is tied to a particular user. Events are applied (and undone) at the level of granularity of dataset, which is defined as a collection of minor events related to a user. From the moment a user is logged into the system, all minor events are stored in the *session* of this user (described above). A dataset can be created by a user from the available events in the current session. Associated with each citation record, there is an initial set of citation metadata. This initial set of metadata may have come from various sources, such as annotation tools, academic search tools or manual insertion through the user interface. The first dataset will be applied to the initial citation metadata. The citation metadata of a record at a specific moment is the result of applying one or more ordered datasets to the initial citation metadata [19].

There are two key issues that require attention during the process of creating a dataset: (a) Events belonging to a dataset must be on the same citation, i.e., we do not allow events related to different citations to be in the same dataset; (b) The order of the event time-stamps is important in that the events of a dataset are applied in the order specified by their time-stamps.

In the current implementation, users can choose any set of consecutive events on a citation to form a dataset. Unless the user defines one or more datasets on the collection of events for a particular user session, all the stored events will be lost when the session ends [19].

3.3 The SRG System Modules

The SRG system consists of several modules. Each module has the same layered design consisting of a client layer, a Web layer, and a data layer. We discuss the technologies and software packages used in the implementation of each module:

The c*lient layer* of each module is composed of Java Server Pages (JSP). The JSP pages communicate with the Web layer over HTTP protocol through SOAP messages.

Table 1. The APIs Used in Implementing the Web Layer

API	Purpose
JDOM	For parsing XML documents
Jakarta Commons HTTP Client	For handling HTTP communication
XPATH	For querying an XML document object
Castor	For XML-to-Java or Java-to-XML binding
JTidy	For parsing HTML documents
Apache Axis	For creating Java Web Services

The *Web layer* is a collection of Web services. The Web services are built using WSDL and SOAP. WSDL is a subset of XML that is used to describe the Web services and their location. SOAP is an XML-based lightweight protocol for exchanging information. The Web service provides methods for communicating with external tools. A number of APIs, summarized in Table 1, are used in the implementation of Web services. Web services are created using Apache Axis. The software modules are deployed in an Apache Tomcat Web container (SRG currently uses Tomcat version 5.0.28).

Web services communicate with the *data layer* using Java Database Connectivity (JDBC). The data layer is composed of several local and remote databases used for storing user specific information, such as the citation records, their access rights, datasets, and so on. Currently, we use MySQL as the Database Management System.

The SRG system consists of the following modules: (A) Session and Event Management; (B) DO Management; (C) Annotation Tools; (D) Search Tools; (E) Authentication and Authorization; (F) Other.

3.3.1 Session and Event Management. This module provides a mechanism for storing data about the current user. These user specific data might be user authentication credentials, any modifications to a DO (called minor events and described in [19]), and the selected "view options" , which control the level of detail with respect to the metadata fields displayed for each DO. Once the user logged in the SRG system, the user's all authentication credentials, all minor events for each DO, and view options for metadata fields of a DO are all maintained in the user session. When a user logs out from the SRG system, all unused minor events (modifications to a DO) for a dataset creation are removed from the current user's session.

3.3.2 DO Management. This module integrates PubsOnline software—"an open source tool for management and presentation of databases of citations via the Web" [21]—into the SRG system and provides an interface for searching the local/remote databases of SRG. It also provides a user with an interface: (1) to manually insert a DO into one of the local/remote SRG databases; (2) to access to the history of a DO, from its entry into SRG system to present; (3) to view detailed information about a DO; (4) to update any metadata fields of a DO, which is saved into session as a minor event for this DO; (5) to perform basic and advanced (a more refined) search in the local/remote databases.

3.3.3 Annotation Tools. This module implements an interface that allows a user to manage the social bookmarking tools: Delicious [22], CiteULike [2], and Connotea [3]. Through this module a user can: (1) upload DOs data and metadata to one of these social bookmarking websites; (2) download DOs data and metadata from one of the social bookmarking websites into one of the local/remote SRG research databases; (3) transfer DOs data and metadata between these social bookmarking websites.

3.3.4 Search Tools. This module allows users to search academic papers and journals through interface using GS and WLA web tools. The search results are displayed as DOs in this module. User can store DO, if they have write access for databases used in the system. However, if selected DO is already existed in selected database, user can update the DO using the session and event management module discussed in Section 3.3.1

3.3.5 Authentication and Authorization. This module provides protection for DOs in SRG system. Users need to be authenticated providing username and password. The system defines three types of users (*Owner, Group, Other*) mentioned in Section 3.2. Users manage their DOs and a database using Read, Write, and Delete access rights.The flexible level of control of authorization is implemented using Super and Group Administrator.

3.3.6 Other. Other than the mentioned modules above, the SRG system has the following other modules: (1) User registration module; (2) Username and password recovery module; (3) User's profile management module, where a user can edit personal information, modify system password, and request subscription to available

SRG system groups; (4) DO metadata "view options" mechanism, which allows a user to define the metadata fields of a DO to be displayed or hidden.

4. Future Work

A desired feature of the system would be a tool that enables harvesting the full text (in PDF, or PS formats) for paper collections defined in various ways, such as the papers of a research group, the papers of a publication venue, and so on.

SRG will develop tools that perform two types of semantic analysis on full-text papers: (a) General metadata extraction, which consists in extracting front-end metadata, such as the title, and authors, and back-end metadata, such as the citations to other papers; (b) Domain specific metadata extraction, which consists in extracting scientific information names from the paper.

5. Conclusion

In this paper we discussed the SRG system, which provides a set of tools and services for supporting scientific research. We described the current state of the development of this system and outlined several direction of the future work.

References

[1] T. O'Reilly, "What is Web 2.0: Design patterns and business models for the next generation of software," 2005.

[2] CiteULike web site. http://www.citeulike.org

[3] Connotea web site. http://www.connotea.org

[4] Bibsonomy web site. http://www.bibsonomy.org

[5] S. Weerawarana, F. Curbera, F. Leymann, T. Storey, and D. F. Ferguson, *Web services platform architecture: SOAP, WSDL, WS-Policy, WS-Addressing, WS-BPEL, WS-Reliable Messaging, and more* Upper Saddle River, NJ Prentice Hall, 2005.

[6] T. Hey and A. E. Trefethen, "Cyberinfrastructure for e-Science," *Science,* vol. 308, pp. 817-821, 2005.

[7] G. Fox, "Collaboration and Community Grids," in *Proceedings International Symposium on Collaborative Technologies and Systems (CTS 2006)* 2006, pp. 419-428.

[8] C. Lee Giles, K. Bollacker, and S. Lawrence, "CiteSeer: An automatic citation indexing system," in *Proceedings 3rd ACM Conference on Digital Libraries (DL'98)*, Pittsburgh, PA, 1998, pp. 89-98.

[9] J. Giles, "Science in the Web age: start your engines," *Nature,* vol. 438, pp. 554-555, 2005.

[10] Semantic Research Grid Project (SRG) web site. http://gf6.ucs.indiana.edu:58080/SRGrid/index.jsp

[11] B. Lund, T. Hammond, M. Flack, and T. Hannay, "Social Bookmarking tools (II): A case study - Connotea," *D-Lib Magazine,* vol. 11, 2005.

[12] G. C. Fox, A. F. Mustacoglu, A. Topcu, and A. Cami, "SRG: A research grid for Cyberinfrastructure based scientific research," Community Grids Lab, Indiana University 2006.

[13] T. Sadeh, "Google Scholar versus Metasearch Systems," in *High Energy Physics Libraries Webzine*. vol. 12, 2006.

[14] P. Jasco, "Windows Live Academic ", 2006. http://projects.ics.hawaii.edu/~jacso/gale/windows-live-acad/windows-live-acad.htm.

[15] R. Tennant, "Is Metasearching dead?," in *Library Journal*, 2005.http://www.libraryjournal.com/article/CA622685.html

[16] B. Quint, "Windows Live Academic Search: The Details," 2006.http://www.infotoday.com/newsbreaks/nb060417-2.shtml

[17] M. K. Bergman, "The Deep Web: Surfacing Hidden Value " in *Journal of Electronic Publishing*. vol. 7, 2001.

[18] C. Anderson, *The long tail: why the future of business is selling less of more*: Hyperion, 2006.

[19] M. Ahmet Fatih, T. Ahmet E., C. Aurel, and F. Geoffrey, "A Novel Event-Based Consistency Model for Supporting Collaborative Cyberinfrastructure Based Scientific Research," in *Collaborative Technologies and Systems CTS 2007* Orlando, 2007.

[20] D. S. Rosenblum and B. Krishnamurthy, "An event-based model of software configuration management," in *Proceedings 3rd International Workshop on Software Configuration Management*, Trondheim, Norway, 1991.

[21] A. M. Scott, K. Richard, L. Matt, and S. Craig, "PubsOnline: open source bibliography database," in *Proceedings of the 33rd annual ACM SIGUCCS conference on User services* Monterey, CA, USA: ACM Press, 2005.

[22] Delicious web site. http://del.icio.us