Web 2.0 for E-Science Environments

Geoffrey C. Fox, Marlon E. Pierce, A. F. Mustacoglu, and Ahmet E. Topcu
gcf@indiana.edu mpierce@cs.indiana.edu
Community Grids Laboratory, Indiana University,
501 North Morton Street, Suite 224
Bloomington, IN 47404

Abstract: We examine the potential impact of Web 2.0 approaches to e-Science and Grid computing. We provide an analysis of Web Service and Grid computing core concepts, which we then map to corresponding concepts in Web 2.0 systems. As we show, Web 2.0, taken collectively, must be viewed as a comprehensive distributed computing approach. We then examine social bookmarking and tagging as an exemplary Web 2.0 service. Tagged bookmarks can be used to build up keyword-based profiles that can be used in collaborator matchmaking services. To be professionally useful to researchers and faculty, these tools need to provide interfaces to scholarly articles for bookmarking. This introduces another level of Web 2.0 service, the Semantic Research Grid, which we overview. We conclude with a discussion of the need for building hybrid Web Service/Web 2.0 systems.

I. Introduction

Distributed computing research to support global scientific challenges in computing, collaboration, and data management ("Grids") have from the beginning been strongly linked to general problems in network-based computing, with the "bag of services" approach of Globus contrasting with the "distributed object" approach of projects such as Legion and CORBA [1].

Today, scalability requirements in Grid computing typically dictate that (at the largest scales) Grids will be loosely coupled collections of nearly stateless network services communicating through well defined, over-the-wire messages. With the introduction of Open Grid Service Architecture concepts and their subsequent refinements [2], Grids have closely tracked Enterprise standards in Web Services, particularly WSDL and SOAP.

Distributed software systems are being revolutionized by developments from e-commerce, e-Science and the consumer Internet. There is rapid progress in technology families termed "Web services", "Grids" and "Web 2.0". The emerging distributed system picture is of distributed services with advertised interfaces but opaque implementations communicating by streams of messages over a variety of protocols. Complete systems are built by combining either services or predefined/pre-existing collections of services together to achieve new capabilities.

The challenges for Grids for the next several years are thus inherited from challenges to Web Services and Enterprise computing generally. Currently, the growth of Web Services is hampered by complexity and lack of broad adoption of major quality of service standards for reliability, security, policy, and workflow composition. Complexity in particular is an important issue, as most Web Service standards have become too complicated to use without vendor- or community-provided development tools.

As we present in this paper, we believe that the next major shift in approaches to Grid architectures is now underway. Just as Enterprise computing approaches are being challenged by so-called Web 2.0 network programming techniques (described below), e-Science architects too will need to examine these approaches. The promise of Web 2.0 for e-Science is that it will provide a much lower entry barrier to developers and enthusiasts, just as the Web 2.0 generally has enabled a very broad range of network programmers with varying degrees of skill.

II. Comparison of Web 2.0 and Grids

The discussion of Grids is confused by many different definitions. One can use the term Grids in narrow fashion to, for example, require use of Web Services or the Web Service Resource Framework or just call any distributed collection of services as "Broad Grids" which is what we do here. Then one uses the term "Narrow Grid" to refer to any "Broad Grid" implemented using particular technology or for a particular application [3]. One very important Narrow Grid is under design by the Open Grid Service Architecture (OGSA) group in Open Grid Forum [4] and another would be the many mashups using Google maps [5]. Our specific goal in this section is to demonstrate that Web 2.0 provides a comprehensive set of "Narrow Grid" implementations of the core "Broad Grid" concepts that are analogous to OGSA and Enterprise Web Service standards.

Web 2.0 is characterized by a suite of important sites and services like YouTube, Google Maps, and Flickr. However it also (currently) uses a distinctive set of technologies including PHP, JavaScript, JSON, AJAX , Microformats and the REST (HTTP) protocol. They feature "Start Page" e.g. (Google Gadgets) user interfaces and construct mashups to build new web applications. We survey these technologies in [12].

Popular technologies for narrow Web Service Grids are Apache Axis, BPEL, WSDL, and SOAP with portlet interfaces and workflows to compose services. Globus and

Condor are important Narrow computing Grid subsystems. Narrow Grid deployments include TeraGrid, EGEE and Enterprise systems from IBM, Oracle, Platform, Univa etc.

Today most e-Science applications are built as Web Service Narrow Grids but we expect growing use of Web 2.0 technologies. Conversely it seems possible that Web 2.0 will turn to Web Service technology when its features such as robustness are required. The approaches agree that systems should be built out of services but make different choices for protocols and API's. Web services use the sophisticated SOAP protocol and WSDL to specify API, which are stored in, for example, UDDI repositories. Web 2.0 typically uses a simple REST invocation pattern that can return XML, including SOAP but more commonly RSS or Atom. Each Web 2.0 service advertises its API as in the nearly 500 at the Programmable Web site [5].

The approaches are not incompatible as both are built of services exchanging messages and these messages can be translated either in specialized mediation service or by libraries at the service endpoints. Although we do not expect Web Services to be as dominant as appeared likely a few years ago, we expect both Web 2.0 and Web Service technologies to be important in the future and systems to take "the best of the best" and integrate together application capabilities and technologies from both Web 2.0 and Web Service arenas. We need to learn how to build such heterogeneous Broad Grids, which will use a "Broad Grid of Narrow Grids" architecture with heterogeneous component Grids using internally some uniform technology federated together.

We see this emerging already. E-Science has tended to use Web Services while adopting a growing number of Web 2.0 goodies like Blogs and Wikis. Web 2.0 storage and computing services like Amazon S3 and EC2 are also growing in popularity. We follow myExperiment's view [6] that one should embrace useful Web 2.0 features and technologies and integrate them with Web Service and OGSA Grids into operational e-Science systems. For example user interface Gadgets have some features lacking in portlets, while some find mashups an easier approach to service composition than Grid workflow.

We summarize a comparison of Web 2.0 and Web Service approaches in tables 1 and 2. Table 1 classifies Web Service capabilities in ten areas using examples; this is taken from Refs [7] and [8], which have more details. Table 2 illustrates how these capabilities are supported in Web 2.0.

Table 1: Ten Web Service Areas with Examples

WS-* Area	Grid/Web Service Examples
1: Core Service Model	XML, WSDL, SOAP
2: Service Internet	WS-Addressing, WS-MessageDelivery; Reliable Messaging WSRM; Efficient Messaging MOTM
3: Notification	WS-Notification, WS-Eventing (Publish-Subscribe)
4: Workflow and Transactions	BPEL, WS-Choreography, WS-Coordination
5: Security	WS-Security, WS-Trust, WS-Federation, SAML, WS-SecureConversation
6: Service Discovery	UDDI, WS-Discovery
7: System Metadata and State	WSRF, WS-MetadataExchange, WS-Context
8: Management	WSDM, WS-Management, WS-Transfer
9: Policy and Agreements	WS-Policy, WS-Agreement
10: Portals and User Interfaces	WSRP (Remote Portlets)

Table 2: Web 2.0 Approach to Web Service Capablities

WS-* Area	Web 2.0 Approach
1: Core Service Model	XML becomes optional but still useful SOAP becomes JSON RSS ATOM WSDL becomes REST with API as GET PUT etc. Axis becomes XmlHttpRequest
2: Service Internet	No special QoS. Use JMS or equivalent?
3: Notification	Hard with HTTP without polling– JMS perhaps?
4: Workflow and Transactions (no Transactions in Web 2.0)	Mashups, Google MapReduce Scripting with PHP JavaScript ….
5: Security	SSL, HTTP Authentication/Authorization, OpenID is Web 2.0 Single Sign on
6: Service Discovery	http://www.programmableweb.com
7: System Metadata and State	Processed by application – no system state – Microformats are a universal metadata approach
8: Management==Interaction	WS-Transfer style Protocols GET PUT etc.
9: Policy and Agreements	Service dependent. Processed by application
10: Portals and User Interfaces	Start Pages, AJAX and Widgets(Netvibes) Gadgets

In addition to these distributed computing concepts, we identify the following analogies between e-Science concepts and Web 2.0.

Table 3: Additional E-Science and Web 2.0 Analogies, I

	Grid/E-Science Implementaiton
1: Community	Designed to enable Virtual

Building	Organizations based on collaborations between existing organizations such as research groups and supercomputing centers. Top-down approach, closely tied to PKI-based security infastructure.
2: Collaboration	Focused on real time audio/video collaborations such as Access Grid. Typically decoupled from Virtual Organizations.
3: Semantic and ontological representation of metadata	Semantic Grid efforts follow closely the Semantic Web and use RDF, OWL for information representation. These can be used for both describing metadata and the contents of digital libraries as well as workflows.

Table 4: Additional E-Science and Web 2.0 Analogies, II

	Grid/E-Science Implementation
1: Community Building	Web 2.0 communities are typically networks of emergent groups with shared interests. Facebook, MySpaces, and Flickr are prominent examples.
2: Collaboration	Dominated by asynchronous collaboration: group-edited content (Wikis), shared commenting/rating/tagging of online content.
3: Semantic and ontological representation of metadata	Metadata described by Microformats (semantic XHTML extensions) that represent community consensus and convention. Ontologies are replaced by "folksonomies" of conventional tags used to describe a network entity.

We now review social bookmarking as a Web 2.0 service exemplar.

III. Tagging, Social Bookmarking, and Profiling

A. Overview

One of the cornerstone concepts of Web 2.0 is its fostering of emergent networks of collaborators. Wikipedia, an online encyclopedia consisting entirely of contributed and voluntarily edited content, is a famous and familiar example of this phenomenon. However, perhaps the more conceptually fundamental example of online community building is shared social bookmarking. Social bookmarking allows users to store their bookmarks of interesting URLs online. These may then be shared with members of one's network or with the general public. Bookmarks are associated with user-provided metadata (comments and keywords, or tags). Prominent examples include del.icio.us and Connotea. Social bookmarking services illustrate several of the core concepts of Web 2.0 services.

Tagging: users provide keyword tags that describe particular URLs that they bookmark. These tags ("web2.0" for example) collectively provide metadata about sites. By adopting (with prompting from the service) conventional tags, online digital entities can have "folksonomic" descriptions built up.

Searching and Discovery: user-supplied tags power effective keyword searching in obvious ways. More interestingly, however, they also enable information-push. Users of the bookmarking service can subscribe to tag feeds (typically formatted in RSS or Atom) in order to receive updates on the latest posts to a particular tag.

Network Building: In the process of tag searching, it is typical that one finds like-minded bookmarkers who are particularly adept at discovering interesting sites. Social bookmarking sites thus typically support RSS subscriptions to tags from a particular person or group of people. These can be open or closed networks.

Multiple User Interfaces: Web 2.0 services may be accessed in numerous ways. We describe this in detail below.

B. Multiple Access Interfaces for Web 2.0 Services

The key architectural feature of social bookmarking sites (and Web 2.0 services in general) is that they support multiple forms of interaction. We generally identify the following four such interfaces.

1. Web interfaces, accessed by users through Web browsers, are of course the most typical.

2. RSS/Atom feeds, as we have already mentioned, allow users get new entries to particular tags pushed to them. Feeds are ubiquitous and easily embedded in browser "start pages" such as MyYahoo, iGoogle, and Netvibes, as well as mobile devices with limited screen real estate.

3. Web programming interfaces, typically developed to support REST-style access (i.e., HTTP PUT and GET), allow developers to work with bookmarking sites programmatically using their own interfaces. Typically the service is invoked by constructing URLs with predefined HTTP request parameters. The response message can be in XML (either RSS/Atom or the site's custom format), Javascript Object Notation (JSON), or any Web microformat. SOAP is also possible but in practice not widely used. Security is handled typically using HTTP authentication. Thus it is possible to uses sites like Connotea as a backend service to one's own service. This is the fundamental concept that enables composite sites (mash-ups).

4. Web Badges and Gadgets: in addition to their main Web sites, many Web 2.0 applications provide exportable Javascript "badges" and "gadgets" that users can embed in other web pages such as Blogs and Start Pages. These may be similar to RSS feeds but provide more formatting. For example, del.icio.us's exportable badge supports tag cloud displays.

E. Final Remarks on Social Bookmarking and Tagging

Before turning to our own work in this field, we believe a brief discussion of some obvious issues is needed. Our focus in this paper is on examining the architecture and application of social bookmarking, which is our research focus. However, we conclude with a few observations and remarks on semantic issues. Tags of course still present fundamental research problems. They are obviously language dependent and can (even in the same language) depend on context. The context problem can be solved in part by examining the tag space, which can be represented as a connected graph with clusters (or cliques). A single tag ("web") may be ambiguous but additional tags ("spider" or "programming") provide the necessary context. We may expect the translation problem to be approximately solved through tag clique-based context as well.

Generally, we may think of these cliques in tag space as providing something akin to the top-down ontologies used in some Semantic Grid research: the caBIG project, for example, provides strictly defined ontologies for cancer research.

III. Tag-Based Researcher Matchmaking and the CITEAM Portal

We now consider some applications of Web 2.0 to scientific research. Grid technologies have typically focused on the problems of securely accessing remote computing and data resources but have not addressed the issues of scientific community building. Typically, scientists excel at networking within their own informal cliques but as a result know more about research taking place across the globe than down the hall or in the next building on campus.

We are motivated to look at this problem from two simple use cases. First, to enable cross-disciplinary science, it is necessary to provide the mechanisms for scientists to discover areas of overlapping research interest. Second, researchers and faculty at colleges and universities supporting under-represented groups (such as Minority Serving Institutions (MSI's) within the United States) need ways to identify their colleagues at other universities.

We are evaluating the use of tagging and bookmarking services as a way to build cliques and social networks to support MSI researchers as part of our NSF CITEAM funded work MSI-CIEC [9]. As we have noted, bookmarking sites such as Connotea have programming interfaces, so it possible to build custom sites on top of this backend service. Users register for the portal and provide keyword tags to describe themselves and their teaching and research interests. This basic user profile is then supplemented continuously by the users' interaction with the system: a user's profile grows and becomes more descriptive as he or she bookmarks and tags research papers, requests for proposals from funding agencies, useful online course material, and other online professional resources. These tags can be searched and subscribed to by other users of the service. We use Connotea as a backend service, since it has less restrictive limits of usage on its Web programming interface than del.icio.us.

One of the gaps in many social bookmarking services that we hope to address in this project is user matchmaking. While it is possible to proactively identify collaborators through tag searching, there is also value in having this information pushed to users as well, and the associated algorithms for determining clique formation and cluster boundaries are classic computer science problems. Our initial plan is to use common tags to help identify users with overlapping interests. This information can then be pushed to users, who can initiate additional contact. Related to this is the "classified ad" approach, in which funding opportunity announcements can be tagged specifically by users who are looking to join a proposal team.

Finally, we note the "interestingness" property used by Flickr has an analogy to matchmaking. Flickr assigns an "interestingness" score to submitted photographs. The details of the algorithm are not published but the general concept is that a photo's interestingness goes up if it is submitted by certain contributors, if it receives high numbers of clicks or comments, if certain prominent users tag it as a favorite, and so on. We hope to explore these ideas for resource tagging and also as an incentive for users to increase their reputation within the system.

IV. Semantic Scholar Research Grid

The CTEAM portal and bookmarking service assumes that bookmarking has professional and not simply recreational value. Research journal search tools such as Google Scholar and Microsoft Live Scholar provide some capabilities for searching for research journals to tag. However, we identify several shortcomings:

- Journal search services currently have a poor hit rate. For example, in informal tests, Google Scholar returned only about 20% of the Community Grids Lab's publications.
- Services must be federated and integrated. We see the need to provide an adaptive architecture that can integrate search results from many existing journal search services simultaneously.
- Search services must be supplemented by user contributions. We need mechanisms for adding the remaining 80% of our lab's publications to online search services.
- Search services should be full Web 2.0 style services, as we have defined them in Section II. Users should have multiple ways of obtaining information: through browsers and programmatically, and through both push and pull mechanisms. Papers need standard metadata descriptions and tags.
- Scientific journal articles evolve over time. Initial drafts evolve into submitted papers, which finally become accepted papers. Likewise, tags and annotations also evolve and may be edited by multiple users. Information may be both lost and gained during this evolution, so we need a way to track versions and revisions.

- More sophisticated organizational models and access control mechanisms are needed. Draft papers and proposals, for example, may be accessible to one's collaborators or to reviewers, but not the general public. This concept can be easily extended to scientific data.

Our research efforts here are collectively called the Semantic Research Grid [10][11]. The system architecture is illustrated in Figure 1. As shown and described in the caption, this system is intended to address the issues listed above.

One interesting feature of the system is the use of events to track changes to documents and metadata, which collectively form a digital. Events are time stamped entities that encapsulate the changes. Thus by replaying events, it is possible to reconstruct the digital entity at any point in its evolution. Since modifications to a digital entity can come from multiple sources, it is possible that illegal operations can occur. For example, a shared entity may have an annotation field deleted by one user while another user is editing the same field. Time stamped events can thus be used to reconcile the system state.

Conclusions and Futures

This paper has presented a comparison and analysis of Web Service and Web 2.0 standards. We conclude that Web 2.0, although not coordinated by standards making bodies, is in practice addressing many of the same core distributed concepts (the "Broad Grid") as Web Services and e-Science activities.

Online social bookmarking and tagging services serve as the simplest illustration of a Web 2.0 service. We examined some of our architectural efforts to build upon these services.

Web 2.0 exchanges sophistication for simplicity and focuses on practicality rather than completeness, which has resulted in scalability of both usage and development. Online community services such as MySpaces and Facebook tout millions of users, far more than any e-Science virtual organization.

It is certainly possible to provide counter examples to Web 2.0 approaches. Security is an obvious shortcoming from the point of view of many research groups and computing centers. We therefore advocate a hybrid approach for e-Science. Web 2.0 style interfaces represent an important outreach opportunity for making scientific results and information available to the public. Scientific mash-ups composed out of Web 2.0 services should be an important way to involve school children, educators, and enthusiasts in scientific endeavors. However, it is not likely that Web 2.0 will replace more complicated distributed computing infrastructure needed by (for example) distributed high energy physics data analysis. Web 2.0 in this case represents a "cell membrane" interface, a simple, controlled communication medium with the outside environment. Within the membrane, more complicated systems may be required.

Determining the boundary between Web 2.0 and more traditional Grid approaches in this hybrid distributed computing world is our next challenge.

This work is supported by the National Science Foundation, Award Number SCI-0537498.

References

[1] I. Foster and C. Kesselman, *The Grid: Blueprint for a New Computing Infrastructure,* 1st Edition, Morgan Kauffman, 1999.

[2] M. P. Atkinson, D. De Roure, A. N. Dunlop, G. Fox, P. Henderson, A. J. G. Hey, N. W. Paton, S. Newhouse, S. Parastatidis, A. E. Trefethen, P. Watson, J. Webber: "Web Service Grids: an evolutionary approach." *Concurrency - Practice and Experience* 17(2-4): 377-389 (2005)

[3] Geoffrey Fox *Grids Challenged by a Web 2.0 and Multicore Sandwich* CCGrid 2007 Conference Rio de Janeiro Brazil May 15 2007 http://grids.ucs.indiana.edu/ptliupages/presentations/CCGridMay15-07.ppt

[4] Open Grid Forum http://www.ogf.org

[5] ProgrammableWeb Web Site http://www.programmableweb.com/

[6] Carole Goble *myExperiment* OGF19 workshop January 29 2007 http://www.semanticgrid.org/OGF/ogf19/OGF19Goble.ppt

[7] Geoffrey Fox, Alex Ho, Shrideep Pallickara, Marlon Pierce, Wenjun Wu *Grids for the GiG and Real Time Simulations* Proceedings of Ninth IEEE International Symposium DS-RT 2005 on Distributed Simulation and Real Time Applications, pages 129-138 IEEE Computer Society, Montreal October 10-12 2005 http://grids.ucs.indiana.edu/ptliupages/publications/gig/DSRTOct05.pdf

[8] Geoffrey Fox, Alex Ho, Marlon Pierce *Appendix on Grid and Web Services* Technical Report October 2005 http://grids.ucs.indiana.edu/ptliupages/publications/gig/DODGridOpportunitiesAppendix.pdf.

[9] Minority Serving Institution Cyberinfrastructure Empowerment Coalition (MSI-CIEC) Project Site: http://www.msi-ciec.org/eduwiki/index.php/Main_Page.

[10] Ahmet E. Topcu , Ahmet Fatih Mustacoglu, Geoffrey Fox , Aurel Cami Integration of Collaborative Information Systems in Web 2.0 Technical report June 27 2007. Available from http://grids.ucs.indiana.edu/ptliupages/publications/Integration_SKG2007Paper.pdf.

[11] A. F. Mustacoglu, A. E. Topcu , A. Cami, G. Fox, "A Novel Event-Based Consistency Model for Supporting Collaborative Cyberinfrastructure Based Scientific Research." *Proceedings of the 2007 International Symposium on Collaborative Technologies and Systems (CTS 2007).*

[12] M. E. Pierce, G. Fox, H. Yuan, and Y. Deng, “Cyberinfrastructure and Web 2.0” in *Proceedings of HPC2006* July 4 2006 Cetraro Italy. Available from http://grids.ucs.indiana.edu/ptliupages/publications/CIWeb20Chapter.doc

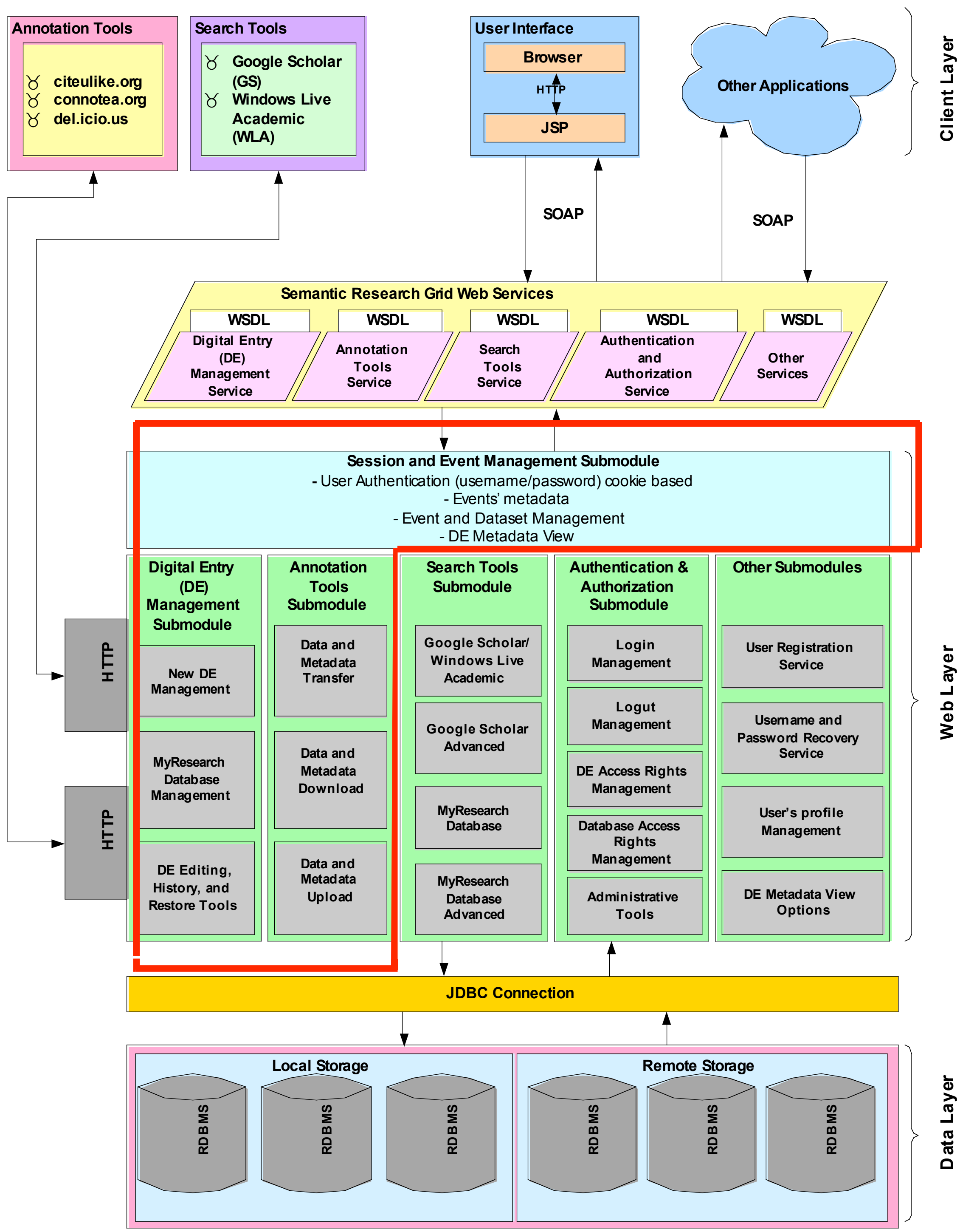

Figure 1 The Semantic Research Grid architecture provides both Web Service and Web 2.0-style interfaces. Pluggable modules provide adapters for accessing Google Scholar and Microsoft Live Search tools as well as contributed databases of articles. Users may interact with the system through portals and services to search, annotate, and collect journal articles. Events are used to track the evolution of digital entities and their metadata.