Linked Data in a Scientific Collaboration Framework

Sudeshna Das, Tom Green, Louis Weitzman,
Alister Lewis-Bowen
Initiative in Innovative Computing, Harvard University
60 Oxford Street, Cambridge, MA USA
01-617-384-5668
sudeshna_das@harvard.edu

Tim Clark
Initiative in Innovative Computing, Harvard University
60 Oxford Street, Cambridge, MA USA
Massachusetts General Hospital, Boston MA USA
01-617-947-7098
tim_clark@harvard.edu

ABSTRACT

In this paper, we describe a Scientific Collaboration Framework (SCF) with semantic underpinnings that is based on the popular content management system Drupal. The framework is designed to support interdisciplinary scientists in publishing, sharing and discussing content such as articles, perspectives, interviews and news items, as well as assert personal biographies and research interests – the basics of any online community. These web materials can then be linked to external knowledge repositories of life science entities such as genes, antibodies, cell-lines or model organisms. Knowledgebases of such important life science entities are being increasingly structured for the Semantic Web either by their publishers or by intermediaries. We have adapted our framework to integrate this emergent semantic content with other types of content in our framework. In this way we organize and repurpose material in online communities by defining and capturing semantic relationships to external knowledge repositories. Such a knowledgebase will enable richer and more powerful interactions amongst many sub-disciplines within the scientific community.

Categories and Subject Descriptors

D.2 [**Software Engineering**]: Software Architectures – *Domain specific architectures*

J.3 [**Computer Applications**]: Life and Medical Sciences – *biology and genetics*

I.2 [**Artificial Intelligence**]: Knowledge Representation, Formalisms and Methods - *Semantic networks*

General Terms

Design, Human Factors, Standardization

Keywords

Scientific collaboration, semantic web, content management system, knowledge integration

World Wide Web Conference'08, April 2008, Beijing, China.

1. INTRODUCTION

In recent times, web-based social networking has the attention of the younger generation and college students especially spend large amounts of time using these sites connecting to one another and exchanging information. One of these sites, Facebook has been shown to be to be used actively by 90% of undergraduate students [8]. While social networking among students is so common, networking among scientific communities is far less common. Bos et al, 2007 [4] describe the barriers to successful communication in their study and list lack of effective knowledge integration, legal and organization barriers and individual style and personalities as some of the important contributors. The opportunistic use of the World Wide Web [2] to effectively build scientific communities and moreover to leverage the Semantic Web [3,9,11] to exchange data in a machine-readable format remains a challenge [15].

That said, there are a few successful communities and one of them is Alzforum www.alzforum.org [12], which is thriving with over 4000 registered users. The site is open to all users but requires registration for participants to comment or take part in discussions. It has a dynamic front page, maintains a database of critical resources and serves as a portal to commonly used resources for the Alzforum research community. The key to the success of the site is proactive solicitation of participation, facilitation and moderation by a team of experts in the field.

In fact, Alzforum has been such a success that other foundations supporting neuroscience research have become interested in creating a "Alzforum clone", but there is no reusable software infra-structure to do so and SCF addresses this particular need.

A project of Alzforum, currently in limited beta-test, is the Semantic Web Application in Neuromedicine, or SWAN [6]. Editorial staff at Alzforum in collaboration with bench scientists use SWAN to capture research claims, hypotheses, evidence and the discourse around them in a semantic format. This format allows the important contradictions and gaps in emergent research to be highlighted and important new research questions to be identified.

An immense opportunity is to create and maintain links between a scientific website's content and other knowledgebases such as SWAN (described above). A few systems for semantic annotation of content have been developed for Semantic Wikis [1,16] but the biomedical community has not yet adopted these. Our framework is designed for compatibility with knowledgebases like SWAN.

Although there is a growing number of biological knowledgebases (GO, CHEBI, SNOMED) [10,7,19] and groups such as the Open Biomedical Ontologies (OBO) consortium

[20], Health Care and Life Sciences Interest Group HCLSIG [17] and other efforts [14] are actively defining common controlled vocabularies and making data available as RDF (Resource Description Framework), the vast majority of biological information continue to reside in isolated laboratory or institution databases. The context of the data is rarely captured and researchers exchange information exchange via emailing of documents or conversations. Community websites publishing online materials rarely, if ever, link them to the biological information or resources, and thus key knowledge is lost. Trivial access to these community resources is critical to successful collaboration well beyond what is possible in individual laboratories or institutions.

We have developed an open-access reusable scientific collaboration framework that can be used by any scientific community to create a website and *link* the content to existing bio-knowledge repositories. The framework draws inspiration from the Alzforum site. It is based on the content management system Drupal (www.drupal.org) and can integrate data from its native relational store and RDF databases. Ease of installation and use is a key feature of the system. The first instance of our software will be used by the *Harvard Stem Cell Institute (HSCI)* to create an open-access site for stem cell researchers. We believe that such a framework is required to achieve optimal productivity and leverage of resources in interdisciplinary scientific research, such as stem-cell research.

2. HARVARD STEM CELL INSTITUTE USE CASE

Harvard Stem Cell Institute (HSCI) is an interfaculty Institute developed to enable and support joint inter-faculty work on stem cell research at Harvard, incorporating the University and all the Harvard-affiliated hospitals. It is presently constituted as a "virtual institute" with central administrative offices and some newly developing core facilities, on a hub-and-spoke model. There are 45 Principal Faculty with an additional 70 Affiliated Faculty across the participating institutions.

The software framework we developed will be used by *HSCI* to create the "*HSCI StemCenter*". A schematic of the sitemap is shown in Figure 1. The site will help *HSCI* laboratories maintain strong current awareness of one another's work; share laboratory methods, resources, and reagents; establish best practices; discover many new opportunities for joint work; create joint work product; and, perhaps most importantly, create a strong community awareness not otherwise easy to foster in the current distributed environment.

As we learned from the successful Alzforum model, the social infrastructure is equally important to foster a successful community. We are working closely with members of the *HSCI* community to define requirements and use cases. *HSCI* has recruited an experienced full-time editor for this project. She has put together an editorial board and together commissioned over 50 articles on stem cell research to prominent researchers in the field. These articles will be linked to related biological entities using SCF; the project goal is to develop *HSCI StemCenter* as an authoritative knowledge source for stem cell research. The articles will be reviewed in the next few months and the site is expected to come online in May 2008.

3. ELEMENTS OF SCF

Our scientific collaboration framework is based on Drupal, an open access content management system. Drupal is built on PHP and uses a relational database to store content. It is very popular, easy to use and administer and is being adopted by more and more communities for bioinformatics needs [13]. Despite its rather modest set of core capabilities, Drupal is highly extensible, with a thriving ecosystem of user contributed modules, and it is in this context that our framework is being developed. As described in Section 3.2, the SCF includes new modules for managing publications, interviews, member information, news items, announcements, and biological entities (e.g., genes). The framework is freely available as a Drupal distribution; however the modules can be used *a la carte* as well. Next, we describe our software design methodology, modules and architecture.

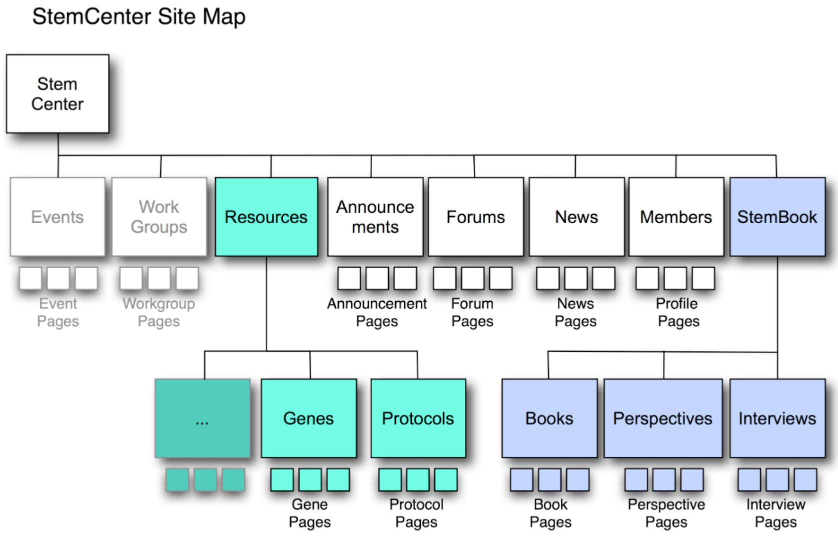


Figure 1. Map of the *HSCI StemCenter site*, an implementation of the Scientific Collaborative Framework

3.1 Design methodology

Our design methodology is composed of four basic phases in an iterative process: an **analysis** phase, a **design** phase, a **prototyping** phase and a **development** phase. The key is to understand the user's goals, tasks and concerns early in the process. We utilize representative personas as a way to measure the usability and usefulness of alternatives. The personas identified for this project include an administrator, a content editor, and a research scientist.

There are many techniques for extracting user feedback, which range from interviews and questionnaires, to low fidelity mockups, to high fidelity prototypes. When building prototypes, there is the trade-off between providing a high degree of detail vs. exploring more alternatives in the given timeframe. Prototypes of both types are important. However, early in the design process, if too much detail is shown, it may detract from the goal of the prototype and focus attention on the wrong aspect of the design. An example of our low-fidelity prototype implementation can be seen in Figures 2.

Through this process we found that in order to become a primary destination for our intended audience (*HSCI*), at a minimum, our site needed to include:

1. A rich source of information via curated publications, interviews and news
2. Access to biological data including genes and protocols, and
3. Groups with appropriate access control

While taking into account the user's needs, we include criteria to evaluate alternatives based on content, navigation and hierarchy, interaction and visual language.

We iterate through analysis, design and prototyping to inform our decisions and help produce alternatives that will survive and flourish through this process.

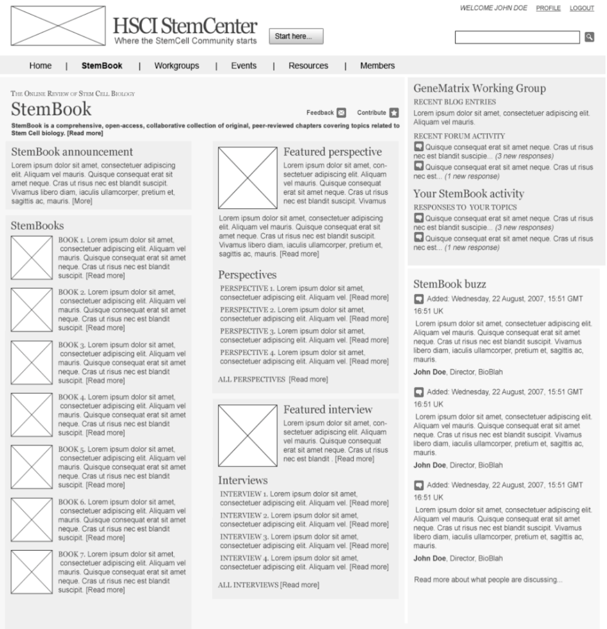

WELCOME JOHN DOE
PROFILE
LOGOUT
HSCI StemCenter
Where the StemCell Community starts
Start here...
Home | StemBook | Workgroups | Events | Resources | Members
The Online Review of Stem Cell Biology
StemBook
Feedback
Contribute
StemBook is a comprehensive, open-access, collaborative collection of original, peer-reviewed chapters covering topics related to Stem Cell biology. [Read more]
StemBook announcement
Featured perspective
StemBooks
Perspectives
ALL PERSPECTIVES [Read more]
Featured interview
Interviews
ALL INTERVIEWS [Read more]
GeneMatrix Working Group
RECENT BLOG ENTRIES
RECENT FORUM ACTIVITY
Your StemBook activity
RESPONSES TO YOUR TOPICS
StemBook buzz
John Doe, Director, BioBlah
Read more about what people are discussing...

Figure 2. Low-fidelity mockup of home page in the StemCenter

3.2 Modules

As a result of the design process, we identified four independent modules for use within the framework. These can also be used independently.

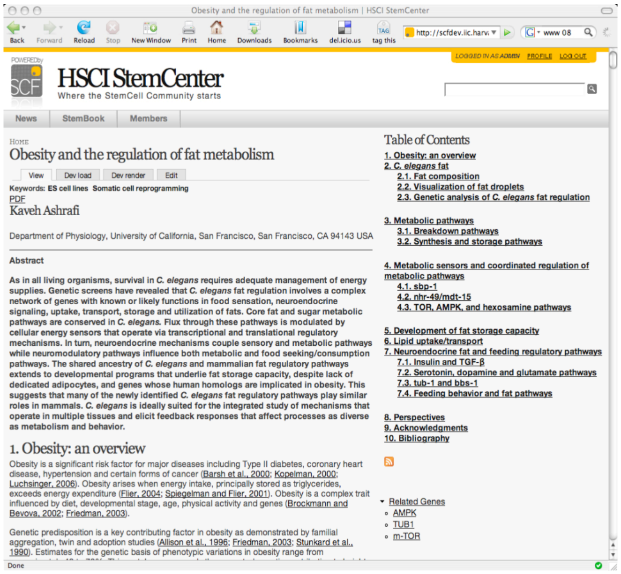
3.2.1 Publications

Publications are the primary content of the site. The framework supports the publication of articles, interviews, news and perspectives. Articles are commonly organized into various books. All publications are implemented as a node type in Drupal and have attributes such as title, author, references, citation and copyright information. This allows us to readily use Drupal features such as comments and tagging.

The articles and perspectives can be created manually or imported from one of these XML (eXtensible Markup Language) sources that support standard publication formats: DocBook (http://docbook.sourceforge.net/) and the NLM (National Library of Medicine) Journal Publishing schemas. XSLT (Extensible Stylesheet Language Transformations) is used to convert the XML for use within Drupal. The software design (discussed in Section 3.3) allows flexibility to import publications available in RDF as well, but till date there has been no driving use case.

The interview module allows an editor to choose participants from the members' list and dynamically add forms for each participant's statements. News items typically include a short teaser and a pointer to some external source where the news has been published.

Once the publications are imported from external sources or created from form-based entry, they can then be linked to related biomedical resources such as genes that are already on the Semantic Web. These related resources or other related publications are also listed on the page. The articles can also be tagged with free text or keywords from a controlled vocabulary as described in Section 3.2.4. Articles are related if they are tagged with the same curated keyword. A screenshot of an example article is shown in Figure 3.



HOME

Obesity and the regulation of fat metabolism

Keywords: ES cell lines Somatic cell reprogramming

PDF

Kaveh Ashrafi

Department of Physiology, University of California, San Francisco, San Francisco, CA 94143 USA

Abstract

As in all living organisms, survival in *C. elegans* requires adequate management of energy supplies. Genetic screens have revealed that *C. elegans* fat regulation involves a complex network of genes with known or likely functions in food sensation, neuroendocrine signaling, uptake, transport, storage and utilization of fats. Core fat and sugar metabolic pathways are conserved in *C. elegans*. Flux through these pathways is modulated by cellular energy sensors that operate via transcriptional and translational regulatory mechanisms. In turn, neuroendocrine mechanisms couple sensory and metabolic pathways while neuromodulatory pathways influence both metabolic and food seeking/consumption pathways. The shared ancestry of *C. elegans* and mammalian fat regulatory pathways extends to developmental programs that underlie fat storage capacity, despite lack of dedicated adipocytes, and genes whose human homologs are implicated in obesity. This suggests that many of the newly identified *C. elegans* fat regulatory pathways play similar roles in mammals. *C. elegans* is ideally suited for the integrated study of mechanisms that operate in multiple tissues and elicit feedback responses that affect processes as diverse as metabolism and behavior.

1. Obesity: an overview

Obesity is a significant risk factor for major diseases including Type II diabetes, coronary heart disease, hypertension and certain forms of cancer (Barsh et al., 2000; Kopelman, 2000; Luchsinger, 2006). Obesity arises when energy intake, principally stored as triglycerides, exceeds energy expenditure (Flier, 2004; Spiegelman and Flier, 2001). Obesity is a complex trait influenced by diet, developmental stage, age, physical activity and genes (Brockmann and Bevova, 2002; Friedman, 2003).

Genetic predisposition is a key contributing factor in obesity as demonstrated by familial aggregation, twin and adoption studies (Allison et al., 1996; Friedman, 2003; Stunkard et al., 1990). Estimates for the genetic basis of phenotypic variations in obesity range from

Table of Contents

Related Genes
- AMPK
- TUB1
- m-TOR

Figure 3. An example article from *HSCI StemCenter*

Any authenticated user of the site can discuss the articles and the discussion relevant threads are displayed at the bottom of each article to draw the reader into the discussion. Users can subscribe to Really Simple Syndication (RSS) feeds for the site to get updates on any content.

3.2.2 Membership Catalog

Members are the essence of any community and it was important to be able to capture the member's contact information, biography and research interests. The research interests are chosen from controlled vocabularies maintained by the site administrator. These were selected from MeSH terms (http://www.nlm.nih.gov/mesh/) and can be easily expanded by the administrator. Such use of controlled vocabulary provides the user an easy access to the expertise of the community and helps establish collaborations. The contributions of a member such as comments or articles are listed on the member page. Members can be searched or browsed by names, affiliations or research interests.

3.2.3 Life Science entities

While it wasn't necessary for publications to be represented in semantic form, various biological entities already have knowledge repositories and the goal was to leverage such content. As a starting point, we have chosen to represent genes, antibodies, cell-lines and model organisms in our system, as these were the most common resources used by the *HSCI* community.

The basic information about the gene record is retrieved from an RDF repository of *Entrez Gene* records [17,18] available at *The Neurocommons* RDF Query interface at http://sparql.neurocommons.org:8890/nsparql/. The architecture and mechanism used to support this query is described in section 3.3. Currently genes from humans, rats, mouse, zebrafish and C. elegans can be retrieved from the repository. Choosing an RDF representation provides a Uniform Resource Identifier (URI) for the gene with which annotation can be asserted. These annotations can be exported as RDF, thus making a SCF based site interoperable with other sites that provide gene annotation. The RDF Query interface also supports complex queries for genes beyond what is possible with an Entrez gene XML repository.

In addition, the site allows users to tag the genes with free tags as well as controlled vocabularies (described in more detail in the following section). Any articles relevant to the community (including the ones published on the site) that discuss the gene are also listed on the gene page. These articles are found using rudimentary text mining.

3.2.4 Community Tagging

We are using the Drupal hierarchical taxonomy system for tagging nodes on the site. We import curated vocabularies from the following sources into our framework:

- Molecular Function, biological process and cellular component - Gene Ontology, GO [10]
- Biomedical terms- Medical Subject Heading, MeSH (http://www.nlm.nih.gov/mesh/)
- Chemical entities – Chemical entities of biological interest, CHEBI [7]

The terms are imported as label-URI pairs, maintaining hierarchical relationships and links to the original definitions. When a user clicks on a term they can access all the nodes that are tagged by that term as well as navigate to the source that describes that term. Type-ahead assistance makes it easy for users to utilize existing terms.

These curated vocabularies provide unambiguous identifiers to the tags. An example usage is shown in Figure 4. An article can be tagged with a GO biological process such as hematopoiesis and a MeSH term such as Zebrafish. Articles published in the HSCI site will have stable Digital Object Identifiers (DOIs) that can be used as unambiguous identifiers. Currently, the arcs between the nodes in SCF and external resources are only of the type 'RelatedTo'. The arc between the GO term and MeSH term provides the context of the relationship and providing such arcs is a future goal of the project. Genes are also annotated in a similar way in SCF. These annotations can also be exported as RDF in SCF. The ability for scientists in a community site to easily create machine-readable annotations on nodes such as articles or genes is the distinguishing feature of our framework.

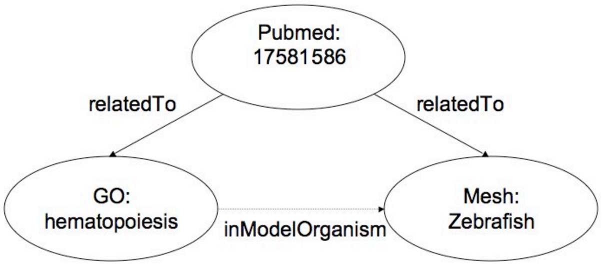


Figure 4: Example of tagging in SCF

We are also using free tagging and later plan to map the resulting folksonomies to existing controlled vocabularies.

3.3 SYSTEM ARCHITECTURE

One of the strengths of the Drupal framework is its extremely rich and flexible programming API. A developer can add arbitrary functionality by creating a new **module** (minimally, a PHP file and attached metadata) and implementing **API** functions in PHP. A site administrator can then install, enable, and configure the module using standard Drupal administration facilities. Because the API is so complete, the actual Drupal kernel is quite small, and essentially all of the standard functionality is implemented in ordinary modules. In addition, there are thousands of freely available user-contributed modules. Often, a site administrator is able to achieve any desired capability for their website simply by installing, enabling and configuring existing modules. For more specialized requirements, though, the programming API makes it possible to modify virtually any aspect of standard or contributed modules by writing new modules in PHP.

The basic unit of content in Drupal is called a **node**. Nodes are classified by **type**, and new types are defined either by dedicated modules or in some cases by configuration of existing modules. There are hundreds of user-contributed modules that either define new node types or contribute additional functionality to existing types. Examples of specific Drupal node types are: story, poll, image asset, bibliography entry and book. Examples of SCF modules are: publication, gene and member catalog.

Despite the abundance of node types, all nodes share a common basic data structure and life cycle API, enabling modules to add functionality which works across all node types. This powerful feature of Drupal means that a node of any type can be commented on, used as a topic for discussion forums, and have taxonomy keywords applied. It may be viewed, edited and deleted, have its revision history tracked and its URL address specified. Sets of nodes can be sorted, searched, used to form syndicated data feeds, and be subject to access control rules. These capabilities and many more are available already, and new node functionality (in the form of contributed modules and enhancements to core modules) is being developed at a rapid rate.

As discussed in section 3.2, the SCF modules define several Drupal node types. Many of these, such as members, interviews and news articles, are intended to be fully authored using the site's web forms. However, other node types, such as genes and antobodies, are based on data that is maintained elsewhere. That is, an SCF gene node is in some sense a "proxy" for a remote gene record. Even so, while a given gene node's actual attribute values may be managed externally to Drupal, we would still like our gene nodes to enjoy all the existing (and future) node-centric capabilities provided by Drupal.

To make the example more concrete, the SCF provides a module called the `gene` module, which defines an ordinary Drupal node type for genes. A gene node contains fields for gene name, symbols, species, summary, and *Entrez Gene* ID, as well as associations between the gene and other biological entity nodes such as model organisms and antibodies. Since there is nothing extraordinary about the `gene` module, gene nodes are by default created like all other normal Drupal nodes, in an editing form where the user is able to enter values directly for all fields.

Gene nodes created manually in this way are perfectly usable in terms of Drupal functionality (i.e. they can be searched, commented on, listed, edited, deleted), but this of course de-normalizes the data. First, many of the field values are taken directly from an external source (in this case the *Entrez Gene* record), and it is tedious and error-prone to enter all this information manually for each gene record we wish to import. Second, the field values would get "stale" over time and need to be updated to track changes in the remote source data. This call for a solution by which simply specifying a remote data source and a unique external identifier (e.g. an *Entrez Gene* ID) would be enough to allow the gene node to serve as a reliable proxy for the remote data.

We considered two ways of accomplishing this. One is to circumvent the data entry and staleness issues by making the stored data representation of gene nodes very impoverished, containing little more than the external ID. The remote data source would then be queried every time the node's field values needed to be displayed. This way, if all data mirroring happens "on the fly" then there are no field values to enter in the form, and no locally cached values to get stale. However, we did not pursue this strategy, partly because performance would undoubtedly suffer without some local caching, but also because it largely negates the value of having genes be first-class Drupal nodes, because many third-party processes, such as sorting, searching, etc., need to inspect stored node attributes in order to do their work. In fact, all nodes are expected to have certain field values (such as title) filled, and this would again require at least some local caching of field values.

The other way to solve the proxying problem is to embrace the fact that some values will be cached locally, and automate the process of copying the remote data into the node's fields and keeping them up to date. At the simplest level, one could imagine a bulk import script which retrieved desired *Entrez Gene* records from the source and created Drupal nodes accordingly. This script could then be run any time the site's users needed to import new gene nodes, and also periodically to keep the information current.

A much more user-friendly solution to these problems would be to augment the `gene` module itself to include the proxying capabilities directly. That is, a gene node would be responsible for querying the external *Entrez Gene* data, and for keeping it up-to-date by periodic update queries behind the scenes. This is a suitable solution but requires building this functionality into the gene module. What if we would like to implement a more general capability, or the same kind of data mirroring for node modules whose source code we do not control?

Our actual solution was to develop a general "node proxying" mechanism, coordinated by a new `nodeproxy` module. `Nodeproxy` does not define any new node types, but rather provides an API for developers to implement type-specific proxying modules that modify the data flow of existing node types. Concretely, we have a `geneproxy` module which uses the `nodeproxy` API to watch for gene-related creation, editing and viewing events, and to inject its own remote data, alter form fields, etc. It is important to note that a proxying module such as `geneproxy` does not define a new "gene proxy" node type, nor is it itself a proxy for gene nodes, rather it provides a way to turn ordinary gene nodes into (partial) proxies for remote data. The Drupal API facilitates this because it allows any module to observe and alter the low-level behavior of other modules (such as database reading and writing, form creation and submission processing).

In the case of `geneproxy`, the remote data is stored in a Semantic Web data store, and the proxying code queries the data via the Neurocommons SPARQL Protocol and RDF Query Language (SPARQL) endpoint, receiving either RDF or SPARQL Query Results XML (http://www.w3.org/TR/rdf-sparql-XMLres/) data in return. This is merely a special case of the more general node proxying capabilities, but we expect it to be the most common and interesting one, and thus we have an `rdfnodeproxy` module, which uses the ARC2 (http://arc.semsol.org/) RDF classes for PHP and provides a common framework above `nodeproxy` for RDF and Semantic Web-backed proxies.

The SCF also uses the proxying mechanism to handle the publishing of scientific articles which are produced externally as XML documents. This allows an SCF site to have publication nodes for these documents without being involved in the creation of their actual content. We currently support article proxying based on compressed file upload (a degenerate case of proxying in which the interaction with the "remote" source requires direct involvement of the user) or by remote HTTP retrieval of previously published XML, such as Creative Commons-licensed Public Library of Science (PLoS) articles which a site editor wants to reproduce locally. This architecture is illustrated in Figure 4.

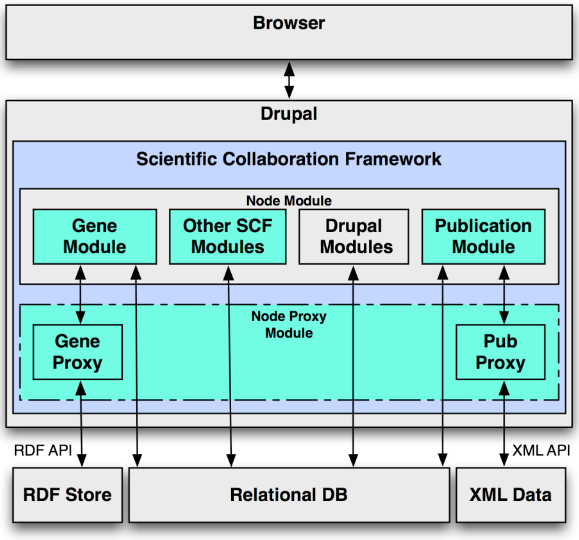


Figure 5. System architecture of the Scientific Collaborative Framework.

Some other features of our `nodeproxy` solution are:

1. a new node proxy module can be applied to any existing node type without modifying the target node module's code or affecting its intrinsic behavior.
2. proxying capabilities can thus be installed, enabled or disabled using existing Drupal module administration features—e.g. the `geneproxy` module itself can be uninstalled, leaving any existing gene nodes untouched; they would still exist but would no longer be proxies, and could then be edited at will.
3. there may be multiple proxy modules targeting the same node type, and interacting in useful ways (e.g. each could be responsible for proxying different sets of fields, or one proxy module might depend on a field value from the other proxy module as the external ID for retrieving its own data)
4. new local nodes may be created in a single click on a remote search result "create-or-view" link, meaning users may never need to see the external ID of a proxy node.

The strategy described here works best when external data records are relatively static and stable. It is currently a read-only implementation, but we will soon be adding both the capability to write back new information to remote services, and also to store new attributes and associations locally, to be made available for external queries. In this way SCF users could enrich the semantic web in a more direct way.

Note that we do not currently provide a general solution for the mapping of data fields from remote formats to Drupal's internal node fields. While the `rdfnodeproxy` module makes this relatively painless for simple import of RDF data, some aspects must still be implemented in the individual node proxying modules. Also, it does not directly support transitive relations to other remote entities. For example, in the Neurocommons data store, each *Entrez Gene* record refers to yet another host species entity, and we do not yet have a general means of automatically connecting this to a local Drupal species node type. In this particular case, the `geneproxy` module contains specific code to map the object of these associations into a simple text "species" field used by the SCF `gene` module.

4. DISCUSSION

While there are a number of efforts ongoing to represent bioinformatics databases as RDF and defining new ontologies for concepts in life–science [17], there are relatively few efforts to integrate and use these existing repositories. Our scientific collaboration framework addresses this critical need and is being used the *HSCI* community and considered by other communities as well (some of these are already in planning stages). Scientific communities can use this framework to create links to life-science entities such as genes and annotate nodes with controlled vocabularies from GO, CHEBI and MESH.

In the future, we plan to integrate with other growing knowledge resources such as SWAN [6]. The controlled vocabularies are currently stored in the Drupal taxonomy system but representing these relations in RDF will enable even more effective knowledge mining and more importantly, interoperability with other sites whose data is represented in RDF. Finally, while manual creation can create wonders such as Wikipedia (http://en.wikipedia.org/) it helps to bootstrap the system with automated text mining. We plan to use text mining to automatically annotate the articles on our site with controlled vocabularies or link them to related resources.

5. ACKNOWLEDGMENTS

We would like to thank Brock Reeve, Executive Director of *HSCI* for his sponsorship, Lisa Girard, Maureen Lyons and members of the *HSCI* communities for extensive feedback on site design and requirements.

We thank Alan Ruttenberg & Jonathan Rees from NeuroCommons for crucial and direction setting advice on site architecture and ontologies.

Last but not least, we would like to thank June Kinoshota and Elizabeth Wu from Alzforum for invaluable consulting on how to create a vibrant scientific community.

6. REFERENCES

[1] Aumueller D. Semantic Authoring and Retrieval within a Wiki. In Proc of 2nd *ESWC*, 2005

[2] Berners-Lee T, Cailliau R, Luotonen A, Nielsen H, and Secret A. The World Wide Web. Communications of the ACM, Vol 37, No. 8, Aug 1994.

[3] Berners-Lee T, Hendler J, Lasilla O. The Semantic Web. Scientific American, May 2001.

[4] Bos N, Zimmerman A, Olson J, Yew J, Yerkie J, Dahl E, et al. From shared databases to communities of practice: A taxonomy of collaboratories. J Computer-Mediated Communication, 2007, 12(2), article 16.

[5] Chabalier J, Mosser J, Burgun A. Integrating biological pathways in disease ontologies.Medinfo. 2007;12(Pt 1):791-5.

[6] Clark T and Kinoshita J. Alzforum and SWAN: the present and future of scientific web communities. Briefings in

Bioinformatics 2007 8(3):163-171; doi:10.1093/bib/bbm012.

[7] Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 2008 Jan; 36(Database issue): D344-50. Epub 2007 Oct 11.

[8] Ellison N, Steinfeld C, and Lampe C. Spatially bounded online social networks and social capital: the role of Facebook. In Proceedings of the Annual Conference of the International Communication Association, 2006.

[9] Feigenbaum L, Herman I, Hongsermeier T, Neumann E and Stephens S. The Semantic Web in Action. Scientific American, December 2007.

[10] Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. Nucleic Acids Res. 2006 Jan 1;34 (Database issue):D322-6.

[11] Hendler J. Science and the Semantic Web. Science 24 January 2003: Vol. 299. no. 5606, pp. 520 – 521 DOI: 10.1126/science.1078874.

[12] Kinoshita J and Strobel G. "Alzheimer Research Forum: A Knowledge Base and e-Community for AD Research," *in* Alzheimer: 100 Years and Beyond. Eds. Jucker M, Beyreuther K, Haass C, Nitsch R, Christen Y. Springer-Verlag, Berlin Heidelberg. 2006:p. 578.

[13] Mooney SD, Baenziger PH. Extensible open source content management systems and frameworks: a solution for many needs of a bioinformatics group. Brief Bioinform. 2008 Jan;9(1):69-74. Epub 2007 Dec 5. Review.

[14] Natale DA, Arighi CN, Barker WC, Blake J, Chang TC, Hu Z, Liu H, Smith B, Wu CH. Framework for a protein ontology. BMC Bioinformatics. 2007 Nov 27;8 Suppl 9:S1.

[15] Neumann E and Prusak L. Knowledge networks in the age of the semantic web. Briefings in Bioinformatics 2007 8(3):141-149; doi:10.1093/bib/bbm013

[16] Oren E. etal. Annotation and Navigation in Semantic Wikis. In *SemWiki WS at ESWC*, 2006

[17] Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, Doherty D, Forsberg K, Gao Y, Kashyap V, Kinoshita J, Luciano J, Marshall MS, Ogbuji C, Rees J, Stephens S, Wong GT, Wu E, Zaccagnini D, Hongsermeier T, Neumann E, Herman I, Cheung KH. Advancing translational research with the Semantic Web. BMC Bioinformatics. 2007 May 9;8 Suppl 3:S2. Review.

[18] Sahoo SS: Converting biological information to the W3C Resource Description Framework (RDF): Experience with Entrez Gene**.** Lister Hill National Center for Biomedical Communications (NLM/NIH); 2006.

[19] Schulz S, Hanser S, Hahn U, Rogers J. The semantics of procedures and diseases in SNOMED CT. Methods Inf Med. 2006;45(4):354-8.

[20] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ; OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007 Nov;25(11):1251-5. Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. Nucleic Acids Res. 2006 Jan 1;34 (Database issue):D322-6.