BRIEFINGS IN BIOINFORMATICS. VOL 10. NO 2. 129–138
Advance Access publication December 6, 2008

doi:10.1093/bib/bbn052

Building biomedical web communities using a semantically aware content management system

Sudeshna Das, Lisa Girard, Tom Green, Louis Weitzman, Alister Lewis-Bowen and Tim Clark

Submitted: 8th August 2008; Received (in revised form): 23rd October 2008

Abstract

Web-based biomedical communities are becoming an increasingly popular vehicle for sharing information amongst researchers and are fast gaining an online presence. However, information organization and exchange in such communities is usually unstructured, rendering interoperability between communities difficult. Furthermore, specialized software to create such communities at low cost—targeted at the specific common information requirements of biomedical researchers—has been largely lacking. At the same time, a growing number of biological knowledge bases and biomedical resources are being structured for the Semantic Web. Several groups are creating reference ontologies for the biomedical domain, actively publishing controlled vocabularies and making data available in Resource Description Framework (RDF) language. We have developed the Science Collaboration Framework (SCF) as a reusable platform for advanced structured online collaboration in biomedical research that leverages these ontologies and RDF resources. SCF supports structured 'Web 2.0' style community discourse amongst researchers, makes heterogeneous data resources available to the collaborating scientist, captures the semantics of the relationship among the resources and structures discourse around the resources. The first instance of the SCF framework is being used to create an open-access online community for stem cell research—StemBook (http://www.stembook.org). We believe that such a framework is required to achieve optimal productivity and leveraging of resources in

Corresponding author. Sudeshna Das, Initiative in Innovative Computing, Harvard University, 60 Oxford Street, Cambridge, MA, USA. Tel: +(617) 384-5668; Fax: +(617) 496-0482; E-mail: sudeshna_das@harvard.edu

Sudeshna Das has been a computational biology researcher for over 11 years. She holds a Bachelor of Technology degree from the Indian Institute of Technology at Kharagpur and a PhD in biomedical engineering from Boston University. Her research interests include design of interoperable bioinformatics systems and analysis of complex biomedical data using statistical and data-mining techniques. She is currently Senior Program Manager at the Initiative in Innovative Computing at Harvard University.

Lisa Girard was trained as a biologist, with a BA from the University of California, San Diego and a PhD from University of California, Berkeley. Dr Girard is trained as a biologist and was involved in the creation, and served as founding editor, of WormBook, an online review of *Caenorhabditis elegans* biology containing nearly 150 original peer-reviewed chapters on topics related to *C. elegans* biology. In July 2007, she joined the Harvard Stem Cell Institute as their Science Editor and has collaborated on the launch, and serves as founding editor of StemBook, an online review of stem cell biology.

Tom Green holds a Bachelor of Science degree in Brain and Cognitive Sciences and a PhD in Linguistics, both from MIT. He has worked as a field linguist with endangered languages in Australia and Central America, and as a software engineer for over 10 years. His specialties are in text processing and relational database application development. He worked as Senior Software Engineer at Harvard Initiative in Innovative Computing and currently he is the group leader for RNAi Informatics at the Broad Institute of MIT and Harvard.

Louis Weitzman has worked at the intersection of design and computers for over 30 years. Louis holds a Bachelor of Architecture from the University of Minnesota, a Master of Architecture in Advanced Studies from MIT's Architecture Machine Group, and a PhD from the Media Lab's Visible Language Workshop. He was a Senior Scientific Software Engineer at Harvard's IIC involved with bringing the design process to emerging projects within the group. He is currently a Staff Engineer at VMware.

Alister Lewis-Bowen has worked on large-scale web site operations and development and in the last 8 years concentrated on bringing user-driven design into agile web application development. Alister worked at the Initiative in Innovative Computing at Harvard University before joining VMware where he is currently working in the Cloud Computing group.

Tim Clark is a researcher in biomedical informatics with over 17 years of experience in the field. He is Director of Informatics at the MassGeneral Institute for Neurodegenerative Disease; an Instructor in Neurology at Harvard Medical School; and a Senior Advisor and Core Member of the Harvard Initiative in Innovative Computing, where he served as Founding Director of Research Programs in 2006–2007. He is also a Founding Editorial Board member of the journal Briefings in Bioinformatics.

interdisciplinary scientific research. We expect it to be particularly beneficial in highly interdisciplinary areas, such as neurodegenerative disease and neurorepair research, as well as having broad utility across the natural sciences.

Keywords: *Semantic Web; scientific communities; ontology; content management system*

INTRODUCTION

Online scientific communities—groups of scientists or collaboratories connected through the Internet—have become an important means by which to exchange data and information. The most common form of an online community is an intra-organization web site. In this format, a department or a lab, for example, shares data and knowledge in a web-based forum.

Barriers to developing a successful community beyond organization or consortia boundaries have been discussed in Bos *et al.* [1]. The obstacles discussed by these authors include issues, such as scientists' preference for working independently and intellectual property competition between institutions. Despite these barriers, the practice of scientists discussing nascent work on the web is an emerging trend, sometimes labeled 'Science 2.0' [2]. Successful scientific communities in which interdisciplinary researchers network and engage in scientific discussions for a common driving cause have been developed and fill a critical resource gap.

One notable example of such a web-based scientific community is Alzforum (www.alzforum.org)—a thriving community of over 4600 researchers networking to find a cure for Alzheimer's [3, 4]. In Alzforum, researchers can discuss papers and news spontaneously and participate in live discussions. Researchers are also invited to provide perspectives on key research news and comment on papers of the week. Compendia of genes, antibodies, animal models and protocols are also available on the site. Currently, the site contains more than 60 000 literature citations, 1900 research news articles, 6000 comments, 20 000 antibodies, 250 research models, 500+ genes from published association studies of late-onset AD, all known mutations causing familial Alzheimer disease, all drugs in Phase 2 and 3 clinical trials and a wealth of community resources, such as databases for grants, conferences and jobs [4]. Another emerging community based on the Alzforum model is the Schizophrenia Forum (www.schizophreniaforum.org)—a community of researchers exchanging ideas to develop better understanding of schizophrenia and improve treatment options.

Communities such as Alzforum and Schizophrenia Forum require both social and technological infrastructure for nurturing their growth [3, 4]. However, both these sites were evolved over time for their specific communities and until recently there has been no common reusable toolkit to create a new site similar in structure. Data and information in these sites and other similar ones are organized and structured in different ways and there was heretofore only limited opportunity to share and exchange information amongst these sites. Moreover, the use of the Semantic Web [5, 6] to exchange information among these scientific communities in a machine-readable format remains a challenge [7].

At the same time, a large number of biological resources are now becoming available as W3C Resource Description Framework (RDF) triples (http://www.w3.org/TR/rdf-primer/). Gene Ontology (GO), CHEBI and SNOMED [8–10] are examples of the most widely used ontologies in the biomedical domain. The ambitious BioMoby project that publishes more than 1400 data sources and analysis tools using a semantic framework has also released its first version [11]. The W3C Health Care and Life Sciences Interest Group [12] and other efforts such as, Open Biomedical Ontologies [13] are actively defining common controlled vocabularies and making data available as RDF. One of the goals of Science Collaboration Framework (SCF) is to annotate the discourse, publications and news published within scientific communities with terms and identifiers from these and other semantically characterized biological information resources, and to make the knowledge and linked data available on the Semantic Web.

There are other efforts to develop collaborative annotation and knowledge management systems using Semantic wikis [14, 15]. Wikis are being increasingly adopted by the biomedical community for collective annotation. Gene Wiki for collective annotation of gene function [16] is a recently published wiki example. Some of these resources are also available as RDF—WikiProteins [17] and BOWiki (http://bowiki.net/). However, wiki is a technology useful for focused annotation efforts and does not easily support community-networking tools such as blogs and forums. Wikis readily enable multiple editing of content and checking the differences between versions. Wiki is a useful technology and is the primary choice when the purpose is to generate a consensus view that is flexible enough to accommodate input from various people. WikiProteins is a great example of that purpose. The entry for human amyloid-β A4 protein precursor (APP), http://www.wikiproteins.org/index.php/Concept:13341741 lists the various functional roles of APP and the types of Alzheimer's disease caused by APP defects. However, the provenance of the claims is lost, and the multi-viewpoints, disagreement or divergence regarding the role of APP in Alzheimer's are not captured in the entry as it is in Alzforum (http://www.alzforum.org/res/for/journal/transcript.asp?LiveID=120). The generalist view of APP presented in WikiProteins is not useful to a scientist specializing in Alzheimer's research. In summary, wiki is most useful for leveraging the 'Long Tail' and synthesizing an encyclopedia from multiple small inputs [16]. In contrast, our framework is aimed at facilitating collaboration, debate and discussion among smaller numbers of specialists in a community.

Indeed, collaboration and networking tools for online communities have also been developed using Semantic Web ontologies and tools. FOAF or 'Friend of a Friend' is the most commonly used ontology in such networks and is used to publish and exchange social information [18]. The Semantically Interlinked Online Communities Project (SIOC—pronounced 'shock') has defined the SIOC ontology, developed technology for interconnecting discussions among communities and allows export of metadata from content management systems into a RDF store [19]. However, biomedical networking and collaboration sites on the semantic web (especially one that annotates published materials with biological resources already on the semantic web) are nonexistent to our knowledge.

We have developed a framework to create interoperable communities through shared ontologies. The SCF is software for building scientific web communities based on a semantically enabled distribution of the open-source content management system Drupal (www.drupal.org). The framework has been designed to leverage existing knowledge repositories and makes use of knowledge and annotation available on the Semantic Web using a subclass of the SWAN ontology [20]. SWAN [3, 20] is an integrated scientific knowledge infrastructure for research communities in neurodegenerative and neuron developmental disorders, enabled by Semantic Web technology and hosted on the Alzforum web community. Discourse among Alzheimer's researchers including hypotheses, claims, evidence and discourse consistency relationships are captured in SWAN. Biological entities such as genes and proteins are linked to discourse elements.

Our framework, SCF, provides a reusable infrastructure to build communities based on the Alzforum principle; these communities using SCF can read and integrate proxied content from 'database-style resources' as well as generate content via the actions of participants and editors. We plan to have this content fed back into SWAN-style knowledge-bases of scientific discourse, either as is or with the additional of further ontology-driven annotation. We believe this approach to developing an interoperable 'community of communities' is extremely promising and we are using it in developing a small ecosystem of communities in the neurodegeneration/neurorepair specialist areas.

SCIENCE COLLABORATION FRAMEWORK

Requirements and design

We developed SCF in an extensive iterative design process based on requirements of an actual community [the Harvard Stem Cell Institute (HSCI)], including development of low fidelity and high fidelity prototypes and user tests [21]. The high fidelity prototypes (Adobe Photoshop documents) were a quick way to get user input without coding investment. The primary person providing feedback was the editor of the site as she was the largest stakeholder in the project. It was difficult for us to get feedback from the time-strapped researchers and we had to rely on just a few users.

We created our first set of requirements from the Alzforum site, but it was quickly evident that the researchers would not be able to devote the time to contribute to the site forums. To maximize participation, we instead decided to publish online review articles on stem cell biology that could be cited by others. The ability to cite was a great motivation for researchers to contribute articles. The next major requirement was to annotate these articles with various biological entities. We had to include text-mining tools to facilitate the otherwise tedious annotation process. Currently, the editor supervises the semi-automated annotation and we hope that over time the community will begin to share the burden as they find utility in the annotation.

When we performed requirements analysis for a second community to be developed on SCF [the Parkinson's researchers community sponsored by the Michael J Fox Foundation (MJFF)], there was a significant overlap with existing features, suggesting that the framework will rapidly reach feature convergence. With a technology like Drupal, only the data structure is defined, branding and theming will always be different for any new site that is developed.

Technology choice

Our goal was to develop software for creating moderated web communities that presents new and challenging research findings that may have divergent interpretations as well as open research problems in a structured manner. To create interoperable communities, we wanted to structure the community knowledge in a machine interpretable format as well as reuse existing, available knowledge bases. It was evident that we needed to make use of the emerging Semantic Web technology for this purpose.

At the same time, for wide adoption, it was important to choose a platform that is accessible to users with various levels of expertise and a Java-based website driven by a RDF triple store did not fit those criteria. So, instead we considered several content management systems, which are typically used to rapidly develop group or institutional sites. We chose to base SCF on a popular one, Drupal, often used by the bioinformatics community [22]. Drupal is a very flexible and modular system and allows the implementation of custom content and functionality by developing a new module, in PHP code, that implements Drupal interfaces. Any community that wants to use SCF and has an existing Drupal site can install a SCF module easily through a forms-based administrative interface. At the other end of the spectrum, a developer can extend SCF functionality and create new custom content. Another advantage of choosing Drupal is that we can leverage a large active community of over 2000 developers. Choice of the latest version of Drupal (6.x), which was in beta when we started our project, had its pros and cons. The latest version included more flexible theming, performance enhancements and support for RDF; however, developing on a version that was not stable was quite challenging.

Software

The SCF consists of the core Drupal software and our custom modules. A community site has the option of installing these modules *a la carte* if there is an existing Drupal site or installing the entire framework. Thus, the site administrator of an existing Drupal website can easily access SCF functionality by installing, enabling and configuring the SCF modules through a forms-based interface. The SCF framework is open source software available under GNU public license at www.ScienceCollaboration.org.

Features

The SCF framework provides the ability to publish articles, interviews and news; annotate these with biological resources such as genes, animal models and antibodies; and create informal discourse of community members around these resources as well as the current scientific articles. The custom modules in SCF to implement this functionality are shown in Table 1. The information of a content type (e.g. gene, article) is contained in a unit called 'node' in Drupal; a site can create instances of gene nodes, article nodes, etc. using these modules. Bioinformatics data tends to be heterogeneous; hence there was a need to instantiate nodes from heterogeneous data sources such as XML and RDF. A flexible architecture needed to be developed for this purpose and is discussed in the next section.

SCF allows representation of various biological resources—genes, animal models and antibodies.

Table 1: Custom modules implemented in SCF

Name	Functionality	Source data
Publication module	Publishing of articles, interviews, perspectives and books	XML
Life science modules	Creation of genes, antibodies and research statements	Input forms and RDF
Member module	Creating a new member profile	Input forms
Taxonomy module	Loading an ontology OBO file into the Drupal taxonomy system	OBO file

The functionality of the modules is listed in the second column. The possible data sources for the modules are listed in the third column.

Our knowledge representation of biological resources is based on, and is an extension of, the SWAN ontology. SWAN-style discourse elements ('research statements') can also be represented in the system using the research statement module. The SWAN research statement is a claim, hypothesis, comment or research question; it may be extracted from a publication or stand on its own [20]. An example graph of a SWAN research statement is shown in Figure 1. The life science entities (three genes in this example) and biological processes from the GO [8] are connected to the SWAN research statement using an instantiation of the 'discusses' relationship. These relationships enrich the statement and are useful in integrating with other statements both within and across communities. If we lookup the Entrez Gene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene) entry for Nanog, a list of general biological processes come up, such as 'transcription' and 'regulation of transcription' which are of less interest to a specialist. Capturing the context with the research statement and making the term more specific makes the (indirect) association of Nanog to 'germ-line cell maintenance' more meaningful.

The data for mammalian genes could have been retrieved from several alternative sources and the most obvious choice would have been the source repository, Entrez Gene. However, the XML data provided by Entrez Gene is not in a machine-interpretable format; hence, and also as a proof of principle of the Semantic Web, we decided to access gene data from a RDF repository. Gene information is imported into the Drupal system from a remote existing Entrez Gene RDF repository [12, 24] using the RDF query language (SPARQL) interface http://sparql.neurocommons.org:8890/nsparql/. The minimal information required to unambiguously refer to the external resource and search for the gene are stored in the Drupal node (similar to the SWAN approach). Currently animal models, antibodies and research statement are defined *de novo* in the system. However, the architecture (described in next section) also supports import of these resources from a remote repository with a SPARQL interface, and this is planned to be the standard approach in later versions of SCF.

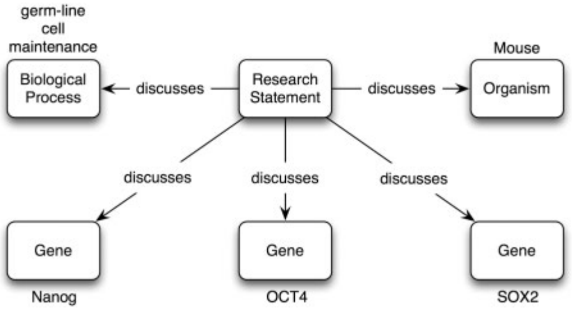


Figure 1: An example graph of a SWAN research statement—'Nanog is thought to function in concert with other factors such as Oct4 and Sox2 to establish ES cell identity' obtained from an article by S. Orkin and colleagues at the HSCI [23]. The genes and biological process are linked to the statement using the 'discusses' relationship.

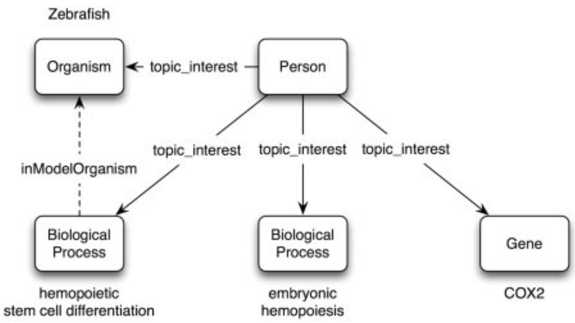
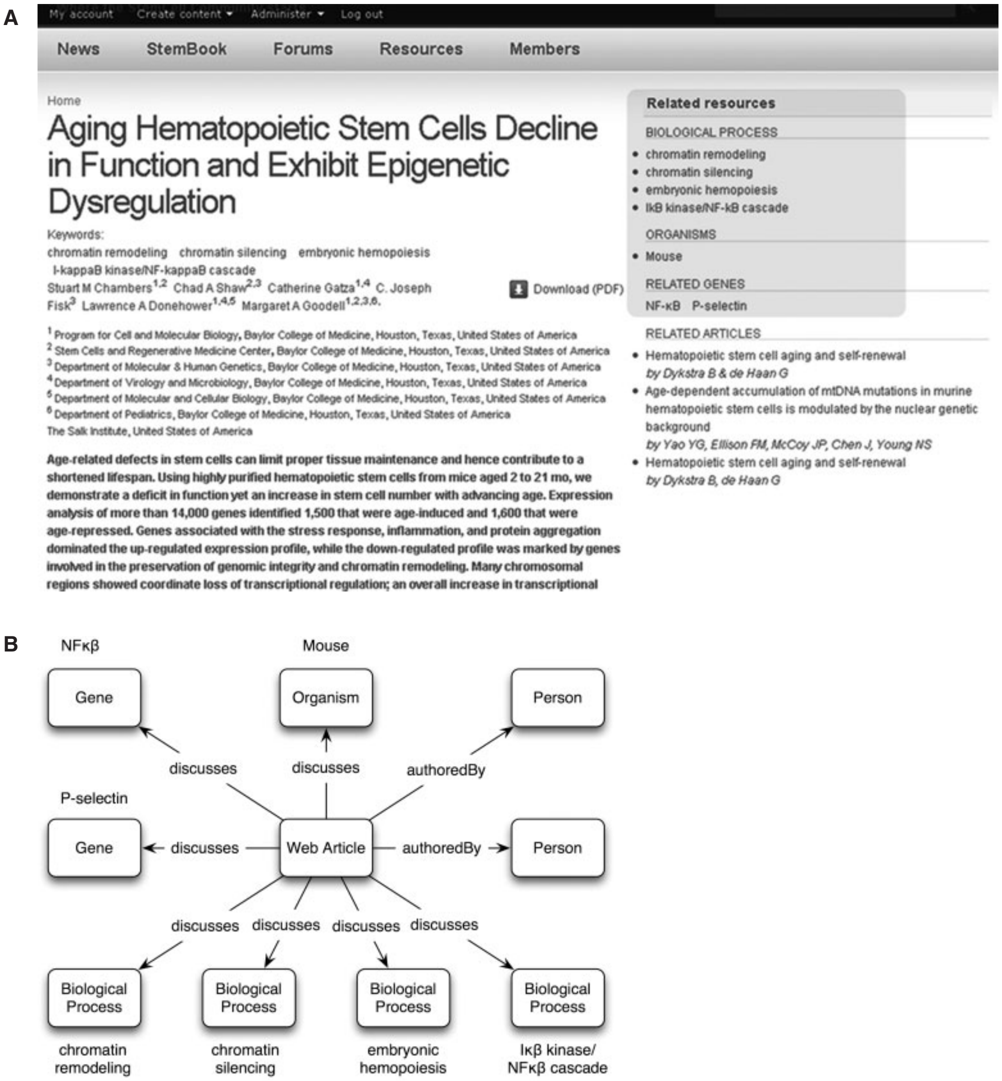


Figure 2: A typical member graph. The person's information is represented by the FOAF ontology. The member's research interests are expressed using an instantiation of the 'foaf : topic_interest' relationship. The relationship 'inModelOrganism' is yet to be developed.

Member information is key to any networking site. Using the framework, members can publish their contact information, affiliation, biography and research interests. Research interests can be picked from any controlled vocabulary that is imported into the Drupal taxonomy system. An example of a member graph that can be exported from the system is shown in Figure 2. We use 'foaf:Person' class to represent the member and 'foaf:topic interest' to capture his or her research interest. The relationship 'inModelOrganism' is yet to be developed and can increase the richness of the represented knowledge.

Articles formatted with XML following a National Library of Medicine Document Type Definition can be uploaded into the site. Articles can be annotated with biological processes, molecular functions or cellular components from the GO. The framework supports loading of GO into the Drupal taxonomy system. The definitions, identifiers and aliases of the GO terms are all imported into the system and the hierarchical relationships maintained. The article can also be annotated with other



Aging Hematopoietic Stem Cells Decline in Function and Exhibit Epigenetic Dysregulation

Keywords:
chromatin remodeling chromatin silencing embryonic hemopoiesis
I-kappaB kinase/NF-kappaB cascade

Stuart M Chambers[1,2] Chad A Shaw[2,3] Catherine Gatza[1,4] C. Joseph Fisk[3] Lawrence A Donehower[1,4,5] Margaret A Goodell[1,2,3,6,]

[1] Program for Cell and Molecular Biology, Baylor College of Medicine, Houston, Texas, United States of America
[2] Stem Cells and Regenerative Medicine Center, Baylor College of Medicine, Houston, Texas, United States of America
[3] Department of Molecular & Human Genetics, Baylor College of Medicine, Houston, Texas, United States of America
[4] Department of Virology and Microbiology, Baylor College of Medicine, Houston, Texas, United States of America
[5] Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, Texas, United States of America
[6] Department of Pediatrics, Baylor College of Medicine, Houston, Texas, United States of America
The Salk Institute, United States of America

Age-related defects in stem cells can limit proper tissue maintenance and hence contribute to a shortened lifespan. Using highly purified hematopoietic stem cells from mice aged 2 to 21 mo, we demonstrate a deficit in function yet an increase in stem cell number with advancing age. Expression analysis of more than 14,000 genes identified 1,500 that were age-induced and 1,600 that were age-repressed. Genes associated with the stress response, inflammation, and protein aggregation dominated the up-regulated expression profile, while the down-regulated profile was marked by genes involved in the preservation of genomic integrity and chromatin remodeling. Many chromosomal regions showed coordinate loss of transcriptional regulation; an overall increase in transcriptional

Related resources

BIOLOGICAL PROCESS
- chromatin remodeling
- chromatin silencing
- embryonic hemopoiesis
- IkB kinase/NF-kB cascade

ORGANISMS
- Mouse

RELATED GENES
NF-κB P-selectin

RELATED ARTICLES
- Hematopoietic stem cell aging and self-renewal
  by Dykstra B & de Haan G
- Age-dependent accumulation of mtDNA mutations in murine hematopoietic stem cells is modulated by the nuclear genetic background
  by Yao YG, Ellison FM, McCoy JP, Chen J, Young NS
- Hematopoietic stem cell aging and self-renewal
  by Dykstra B, de Haan G

Figure 3: (**A**) An article published in PLoS Biology by Chambers *et al*. [25] and imported into SCF. The related keywords including biological processes, cellular components and molecular functions are listed on the right and highlighted. (**B**) The graph of information contained within the article in (**A**). The relationships 'discusses' and 'authoredBy' are part of the SWAN ontology.

biological resources and discourse elements, such as research statements and antibodies. The screenshot of an example article imported into SCF is shown in Figure 3A and lists the related resources in the beginning of the article on the right column. The graph representing this information is shown in Figure 3B. Such a graph can be very useful in determining a person's research interests—one can lookup the most frequent term(s) associated with a person via his authored publications.

Architecture

The SCF framework utilizes a 'Node Proxy' architecture (Figure 4) for selected Drupal nodes defined as resources that are available from the Semantic Web or other web services. The Node proxy module (shown in Figure 4 with dashed lines) provides an API for developers to implement specific proxy modules. The two proxy modules implemented so far are gene proxy and pub proxy modules. The gene proxy module's function is to populate the information of a gene node. It does so from a RDF store via a SPARQL endpoint. If the gene proxy module is not enabled in a site, gene nodes still exist and they can be generated using form based inputs. The pub proxy module populates a publication node from an XML document. The other modules currently do not use the proxy mechanism and depend on obtaining their information from form-based inputs.

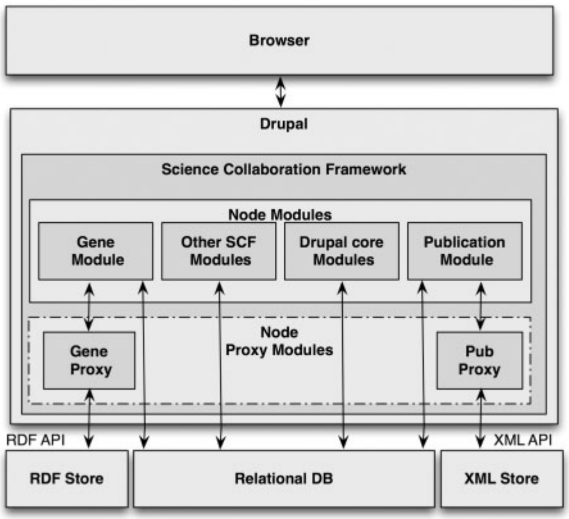


Figure 4: The SCF framework illustrating the 'Node Proxy' architecture.

A detailed description of SCF architecture and functionality can be found in Das *et al.* [21]. This architecture will make it possible in future to define common schemas in OWL for a set of web communities and to enable interoperability across biological resources, SWAN research statements or other objects of interest defined in the shared schemas. We plan to make these graphs discussed above available via RDFa (http://www.w3.org/TR/xhtml-rdfa-primer/) embedded within the HTML and that work is in exploratory stages. We are also exploring ways to extract the knowledge and export it to a RDF triple store.

SCIENTIFIC COMMUNITIES WITH SCF

The SCF framework is first being used by the HSCI to create StemBook—an open community for stem cell biologists. Another SCF-based project with a completed design and currently in the programming stage is PD Online Research—an online community of Parkinson's researchers. Other communities of neurodegenerative researchers are also actively evaluating SCF for use. While each of these communities has a different focus, there is an obvious advantage to basing them on common infrastructure, shared ontologies and sharable modules. As forums for discourse among community members, which can be of significant interest and importance across related specialties, information sharing is an important feature. This can take the place of common repositories of resources and/or discourse. Discourse from any of these communities will—as planned in the overall design—be fed into one or several SWAN knowledgebase(s) allowing for a community of interoperable communities. StemBook, PD Online and their goals are described below.

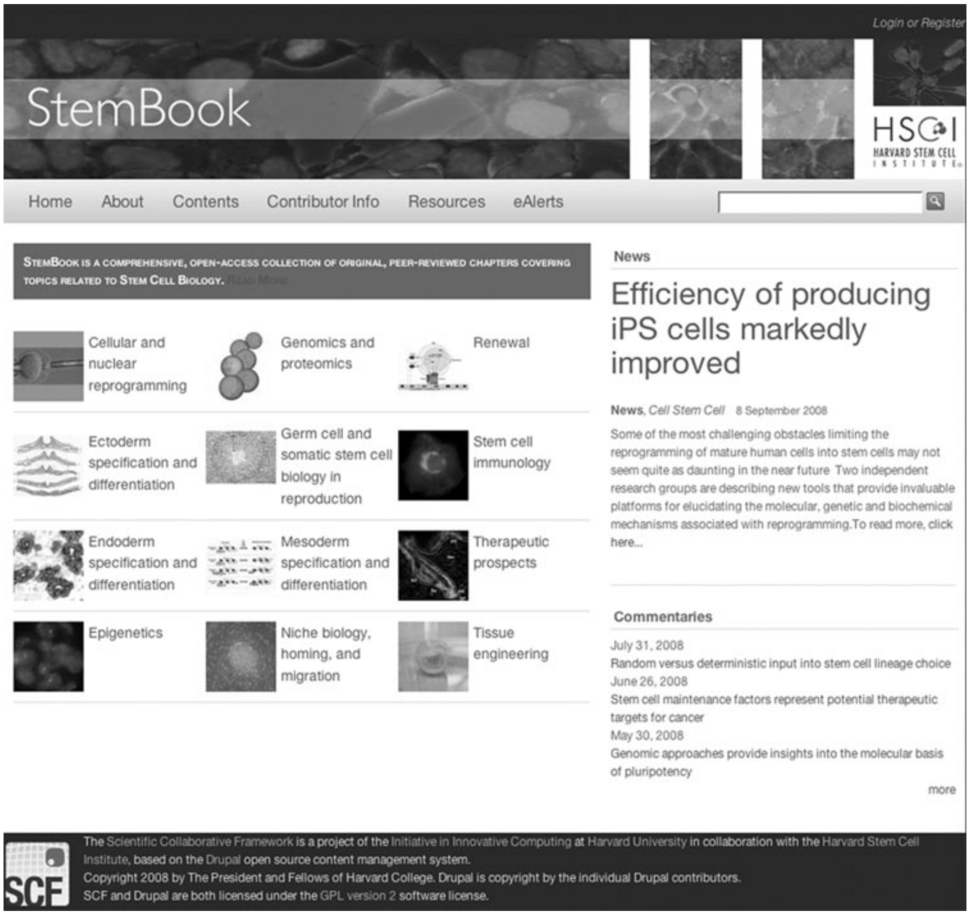
Stembook

StemBook (http://www.stembook.org), is a collection of original, open-access, peer-reviewed chapters covering a range of topics related to stem cell biology written by top researchers in the field at the HSCI and worldwide. StemBook is aimed at stem cell and nonspecialist researchers. It may also be incorporated into undergraduate and graduate curriculum.

StemBook (Figure 5) is divided into sections covering a range of topics related to stem cell biology for example, the niche, homing and migration, endoderm specification and therapeutic prospects. Thus, StemBook will contain relevant, up-to-date information related to stem cell biology that spans the basic to the translational with an infrastructure equipped to continue evolving along with the field it covers.

Currently over 80 StemBook chapters are commissioned, with submission deadlines extending through fall 2009. Nearly 40 chapters have already been submitted and are at various stages of the review and revision process. Every chapter in StemBook represents a citeable source, each receiving its own Digital Object identifier (DOI) and is intended to be indexed in PubMed and included on the NCBI Bookshelf. StemBook is classified as an online periodical and has its own ISSN [international standard serial (periodical) number]. The online format of StemBook is highly amenable to the rapidly advancing field of stem cell biology. Unlike a print release, StemBook can be kept up to date easily. Additionally, the online format supports a range of media, including movies, which would not be possible in a print version.

StemBook content is linked to other relevant resources, such as gene and protein pages using the SCF capabilities, thereby expanding its utility to the reader. And it will support, through built-in functions in the SCF framework, development of robust community discourse amongst a base of registered members. Significantly, the content provides a valuable source of high-quality biological information that can be parsed for semantic applications.


Login or Register
StemBook
HSCI
HARVARD STEM CELL INSTITUTE
Home
About
Contents
Contributor Info
Resources
eAlerts
Cellular and nuclear reprogramming
Genomics and proteomics
Renewal
Ectoderm specification and differentiation
Germ cell and somatic stem cell biology in reproduction
Stem cell immunology
Endoderm specification and differentiation
Mesoderm specification and differentiation
Therapeutic prospects
Epigenetics
Niche biology, homing, and migration
Tissue engineering
more
SCF

StemBook is a comprehensive, open-access collection of original, peer-reviewed chapters covering topics related to Stem Cell Biology. Read More

News

Efficiency of producing iPS cells markedly improved

News, *Cell Stem Cell* 8 September 2008

Some of the most challenging obstacles limiting the reprogramming of mature human cells into stem cells may not seem quite as daunting in the near future Two independent research groups are describing new tools that provide invaluable platforms for elucidating the molecular, genetic and biochemical mechanisms associated with reprogramming.To read more, click here...

Commentaries

July 31, 2008
Random versus deterministic input into stem cell lineage choice
June 26, 2008
Stem cell maintenance factors represent potential therapeutic targets for cancer
May 30, 2008
Genomic approaches provide insights into the molecular basis of pluripotency

The Scientific Collaborative Framework is a project of the Initiative in Innovative Computing at Harvard University in collaboration with the Harvard Stem Cell Institute, based on the Drupal open source content management system.
Copyright 2008 by The President and Fellows of Harvard College. Drupal is copyright by the individual Drupal contributors.
SCF and Drupal are both licensed under the GPL version 2 software license.

Figure 5: The front page of the HSCI StemBook.

PD Online

PD Online Research is a project of our group at MGH in collaboration with the MJFF. Its goal is to develop a large-scale self-organizing community of basic and clinical scientists, industry professionals, grant-makers, philanthropists and financial investors dedicated to making hard decisions about how to spend limited public and private funds to advance the treatment, prevention and cure of Parkinson's disease.

This community is intended to conduct global public workshops for critical analysis of PD research, and identification of new research questions/directions. It will also connect priority research with financial, industry and academic partners.

We are developing a sophisticated community knowledgebase infrastructure on the web to capture this online discourse, integrate it with public biological databases and make it readily searchable and freely available. Privacy management will allow for user-controlled confidentiality of some workspaces. Initial design and user testing is complete and implementation is ongoing.

PD Online is being developed on the SCF platform. PD Online hopes that giving scientists the means, and therefore the responsibility, to directly influence the generation of targeted funding programs will add sharper motivation and drive to these discussions. Again such discourse can fed into a SWAN knowledgebase that will be interoperable with the StemBook knowledgebase.

DISCUSSIONS AND FUTURE WORK

Despite the promises of Web 3.0 to create machine-readable knowledge resources, the vast majority of web content continues to be on the 'traditional' web [18]. A primary reason for this delay in paradigm shift is that Semantic Web technologies require significant technical expertise and there are very few ready-to-use toolkits exist as of today. Our SCF is based on

Drupal, an easy to adopt content management system, and we expect it to lower the barrier to entry for any group to build an online community with a presence on the Semantic Web. Drupal effectively hides the complexity of elements of the Semantic Web from the end user while delivering many of its benefits.

Semantic wikis [14, 15] are another approach for providing semantic content on the web. However, these are most useful for organizing focused annotation for a concept, and are particularly helpful for evolving toward convergence of views [17]. Our framework, on the other hand, is able to provide clear attribution and credits to authors of distinct viewpoints. StemBook, for example, currently comprises approximately 20 articles on different topics that have been through an editorial-review process, and represents the contributions of about 40 eminent stem cell researchers. Each article clearly states the authors and can be cited by other authors. Annotation is done only by the editors of StemBook, and is facilitated by text-mining scripts. Alzforum, a much more established community, has more frequent contributors; they publish 10–15 commentaries a week from thought leaders in the field who have been editorially vetted. A team of editors moderates all comments and contributions on the site. Thus, the invited contributions and moderation presents a different model of a scientific community than a wiki—and, we believe, one that is well adapted to emergent 'hot' areas of research where there is not yet a theoretical consensus.

The online publication of stem cell articles in StemBook with embedded multimedia elements, annotation with biological resources, ability to interact and have an online discussion represents a new trend in publishing. The journal 'PLoS ONE' (http://www.plosone.org/) has similar interactive capabilities and 'Nature network' (http://network.nature.com/) has also developed an active scientific community. We believe that in future, most journals will have a rapidly growing online, interactive component. In an interview published in the Harvard Gazette [26], David Schaffer, professor of chemical engineering at the University of California, Berkeley, and author of one of the chapters in StemBook, says that it fills a niche in an emerging field that has traditionally been filled by textbooks and printed journals.

The success of SCF enabled communities will ultimately be measured by usage statistics such as number and frequency of contributions, page hits and citations in other work. We do not have such statistics for StemBook and PD Online because they are in early stages of development; however, we have preliminary evidence that the effort will be successful from favorable user comments, such as the one described in the earlier paragraph.

The SCF architecture is fully compatible with the Semantic Scientific Community model described in Zhang *et al*. 2008 [27]. As more semantic web resources are robustly implemented, they can be reflected and proxied into the SCF, and related annotations can be exported. This is a significant future project but we believe the way forward can be navigated through incremental development. Another related effort, the SIOC project [19] provides semantics to a community but does not address the need for annotation with biological resources. There is an on-going collaboration between SIOC and SWAN teams to align the ontologies.

We have released SCF version 1.0 with the features discussed in this article. The use of SWAN and FOAF ontology to express knowledge in the examples above is proof of concept and much work remains to fully integrate these ontologies as well as the SIOC ontology within SCF. We have not developed any ontology as part of our project—our goal has been to use the existing ones. The work to export the knowledge is also in preliminary stages and we need to expand ways to access the knowledge base. SCF is currently implemented as part of the Drupal framework and in future, we would like to explore ways to make it work with other Web 2.0 technologies.

In future, we would like to represent additional resources such as cell-lines, laboratory protocols and sequences. We would like to improve our text-mining tools to automate the annotation of discourse with resources. Curators will review these text-mining outputs. Such a resource can serve as a training set for machine learning algorithms to further improve the rules used by the text-mining algorithm.

Key Points

- We are developing biomedical web communities that provide a semantic context of discourse through linked data and ontologies.
- Our SCF framework can create such communities with an easy to adopt content management system.
- We are extending the framework to create communities that can interoperate with each other providing a fertile ground for knowledge discovery.

Acknowledgements

The authors would like to thank Brock Reeve at the Harvard Stem Cell Institute and Michael Rogan at the Michael J. Fox Foundation for their support; Ben Melancon on his work on SCF; Paolo Ciccarese, Alan Ruttenberg and Jonathan Rees for their consulting on Semantic Web technologies; Rosalind Reid for helpful comments on manuscript and finally June Kinoshita and Elizabeth Wu for their ideas and suggestions on scientific community development.

FUNDING

This work was supported by grants from an anonymous foundation; Provost's office at Harvard University, Cambridge, MA; Michael J. Fox Foundation, New York.

References

1. Bos N, Zimmerman A, Olson J, *et al*. From shared databases to communities of practice: a taxonomy of collaboratories. *J Comput Mediated Commun* 2007;**12**(2):652–72.
2. Waldrop MM. Science 2.0. *Sci Am* 2008;**298**(5):68–73.
3. Clark T, Kinoshita J. Alzforum and SWAN: the present and future of scientific web communities. *Brief Bioinform* 2007; **8**(3):163–71.
4. Kinoshita J, Clark T. Alzforum. *Methods Mol Biol* 2007;**401**: 365–81.
5. Berners-Lee T, Cailliau R, Luotonen A, *et al*. The World Wide Web. *Commun ACM* 1994;**37**(8):76–82.
6. Berners-Lee T, Hendler J, Lasilla O. The Semantic Web. *Sci Am* 2001;**284**(5):34–43.
7. Neumann E, Prusak L. Knowledge networks in the age of the semantic web. *Brief Bioinform* 2007;**8**(3):141–9.
8. Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res* 2006;**34**:D322–6.
9. Degtyarenko K, de Matos P, Ennis M, *et al*. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* 2008;**36**:D344–50.
10. Schulz S, Hanser S, Hahn U, *et al*. The semantics of procedures and diseases in SNOMED CT. *Methods Inf Med* 2006;**45**(4):354–8.
11. BioMoby Consortium, Wilkinson MD, Senger M, *et al*. Interoperability with Moby 1.0—it's better than sharing your toothbrush! *Brief Bioinform* 2008;**9**(3):220–31.
12. Ruttenberg A, Clark T, Bug W, *et al*. Advancing translational research with the Semantic Web. *BMC Bioinformatics* 2007;**8**(Suppl 3):S2.
13. Smith B, Ashburner M, Rosse C, *et al*. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;**25**(11): 1251–5.
14. Aumueller D. Semantic authoring and retrieval within a Wiki. In: *Proceedings of 2nd ESWC*, 2005.
15. Oren E, R Delbru, K Moller, *et al*. Annotation and Navigation in Semantic Wikis. In: *SemWiki WS at ESWC*, 2006.
16. Huss JW, Orozco C, Goodale J, *et al*. A Gene Wiki for community annotation of gene function. *PLoS Biol* 2008; **6**(7):e175.
17. Mons B, Ashburner M, Chichester C, *et al*. Calling on a million minds for community annotation in WikiProteins. *Genome Biol* 2008;**9**(5):R89.
18. Finin T, Ding L, Zhou L, *et al*. *Social Networking on the Semantic Web*. The Learning Organization, Vol. 12, No. 5, Emerald Group Publishing, 2005.
19. Breslin JG, Harth A, Bojars U, *et al*. towards semantically interlinked online communities. In: *2nd European Semantic Web Conference*, May 29 to June 1, 2005, pp. 500–514. Heraklion, Greece.
20. Ciccarese P, Wu E, Wong G, *et al*. The SWAN biomedical discourse ontology. *J Biomed Inform* 2008;**41**(5): 739–51.
21. Das S, Green T, Weitzman L, *et al*. Linked data in a scientific collaboration framework. In: *The 17th International World Wide Web Conference*, 2008, Beijing, China.
22. Mooney SD, Baenziger PH. Extensible open source content management systems and frameworks: a solution for many needs of a bioinformatics group. *Brief Bioinform* 2008;**9**(1):69–74.
23. Wang J, Rao S, Chu J, *et al*. A protein interaction network for pluripotency of embryonic stem cells. *Nature* 2006; **444**(7117):364–8.
24. 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.
25. Chambers SM, Shaw CA, Gatza C, *et al*. Aging hematopoietic stem cells decline in function and exhibit epigenetic dysregulation. *PLoS Biol* **5**(8):e201.
26. HSCI creates Web presence for research. *Harvard Gazette* 2008. http://www.news.harvard.edu/gazette/2008/10.09/11-stembook.html (October 2008, date last accessed).
27. Zhang Z, Cheung KH, Townsend JP. Bringing Web 2.0 to bioinformatics. *Brief Bioinform*. (October 2008, date last accessed).