Science Collaboration Framework
a reusable semantic toolkit for scientific communities

Launched 2007
Status Continuing
Participants
Paolo Ciccarese, HMS/MGH
Tim Clark, HMS/IIC/MGH
Stephane Corlosquet, National University of Ireland
Sudeshna Das, IIC
Rahul Dave, IIC/Astronomy
Lisa Girard, HSCI
Mark Goetz, IIC intern
Tom Green, IIC
Alister Lewis-Bowen, IIC
Greg Rundlett, IIC
Louis Weitzman, IIC

Selected accomplishments
- Released the Science Collaboration Framework, a reusable toolkit for scientific communities, in June 2008
- Launched StemBook, online review of stem cell biology powered by SCF, in September 2008
- Created website for SCF project from with software downloads, documentation and support forums
- Completed design and requirements for PDOnline, online community of Parkinson's Disease researchers sponsored by the Michael J. Fox Foundation
- PDOnline beta release, April 2009
We have developed the Science Collaboration Framework (SCF), a lightweight software framework that scientific communities can use to create open-access, online, scientific publications and to effectively share and search common knowledgebase resources on the web. The software uses Web 3.0 technologies (social web, semantic web, text mining) and thus allows interoperability with other Web 3.0 sites. The software allows communities to publish complex scientific articles, annotate them with controlled vocabularies or ontologies, register research interests of members and conduct discussion forums. The software can integrate with other knowledge repositories and the site knowledge is available as linked data. The software is modular, so different communities can install and enable different features as well as contribute modules back to the main framework, thus creating a software community as well.

The first site based on our software, StemBook (www.stembook.org), an online open access peer-reviewed collection of invited review chapters covering a range of topics related to stem cell biology, went “live” in September 2008. Several other sites are under development, including a new web community for Parkinson’s Disease researchers, PD Online, and a re-engineered version of the popular Alzheimer Disease research community Alzforum (www.alzforum.org). Several other, smaller research communities in Europe and United States are also adopting SCF. This project has made a high impact and now has national and international collaborators. The project has both software and a user community. Because the software is open source, other developers have made modifications and bug fixes and contributed back to the framework. SCF has a large and growing user community, thus enabling a community of communities.

The sites developed on the SCF platform are interoperable with each other and with other sites on the Semantic Web. In this new paradigm, there is a significant reduction in artificial barriers between research disciplines, and a much more dynamic and agile approach to information exchange.

**Project goals**
The goal of the SCF project has been to develop a reusable framework that can be used by other biomedical communities and also adapted for use by any scientific community. The framework will allow interdisciplinary scientists:

- to publish and discuss on-line content such as articles, news, and perspectives;
- to share, annotate, archive and retrieve digital content;

Web-based biomedical communities are an increasingly popular vehicle.... However, information organization and exchange is usually unstructured, rendering interoperability difficult.

<table>
<thead>
<tr>
<th>date</th>
<th>milestone</th>
</tr>
</thead>
<tbody>
<tr>
<td>March 2007</td>
<td>Project inception</td>
</tr>
<tr>
<td>May 2007</td>
<td>Received grant from a major foundation, matched by Provost</td>
</tr>
<tr>
<td>June 2007</td>
<td>Initial demo at Harvard Stem Cell Institute retreat</td>
</tr>
<tr>
<td>September 2007</td>
<td>Hired Thomas Green as software engineer</td>
</tr>
<tr>
<td>April 2008</td>
<td>Received grant from Michael J. Fox Foundation for Parkinson disease research community, PDOnline</td>
</tr>
<tr>
<td>May 2008</td>
<td>First release of SCF</td>
</tr>
<tr>
<td>July 2008</td>
<td>Completed requirements for PDOnline</td>
</tr>
<tr>
<td>September 2008</td>
<td>Launched StemBook—online review of stem cell biology</td>
</tr>
<tr>
<td>April 2009</td>
<td>Released PDOnline Beta</td>
</tr>
</tbody>
</table>
• to search, browse and annotate joint resources such as laboratory reagents and protocols; and
• to provide semantic context to on-line content for more powerful interactions within multiple sub-disciplines.

SCF’s development phase launched rapidly in the spring of 2007, when major foundation funding was secured and Sudeshna Das was brought to IIC to manage the project. A demonstration at a retreat sponsored by the Harvard Stem Cell Institute in June 2007 helped define the first phase of the project, which became known as StemBook. Software engineering went into full swing over the next few months as Thomas Green was hired as a full-time software engineer, and the new IIC Design Group, Louis Weitzman and Alister Lewis-Bowen, were tasked with developing the StemBook design.

Research goals
Web-based biomedical communities are an increasingly popular vehicle for sharing information among researchers and are fast gaining an online presence. However, information organization and exchange is usually unstructured in such communities, 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 as Resource Description Framework (RDF).

We have developed SCF as a reusable platform for advanced structured online collaboration in biomedical research. The software allows communities to publish complex scientific articles, annotate them with controlled vocabularies or ontologies, register research interests of members and conduct discussion forums.

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. 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 Science Collaboration Framework fulfills that niche.

It 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.

The SCF architecture is fully compatible with the Semantic Scientific Community (SSC) model (Zhang, Z., K. H. Cheung and J. P. Townsend, Bringing Web 2.0 to bioinformatics. Brief. Bioinform., Oct. 2008). As more semantic web resources are

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.
I began my internship with the IIC just after graduating from high school and prior to beginning my freshman year at Harvard. I had developed an interest in computer science over the course of my junior and senior years at high school. I took two introductory CS courses in Java for the AP test and began teaching myself the basics of web programming outside of school. Already planning on concentrating in computer science, I thought an internship with a Harvard group would be the perfect way to increase my knowledge of the subject and get some valuable experience. I applied for work helping the IIC's Design Group rebuild their website using a tool called Drupal (which I knew very little about at the time) and was accepted.

During my first few meetings with the leaders of the Design Group, Alister and Louie, I learned that I would be helping people within the IIC edit pages for their research projects, submit new content and collaborate online. After I had gotten my bearings and completed a few small tasks for the site, I was tossed into the deep end: given much control of the new website and required to solve problems generally on my own. While I had much work in front of me, I found that the design practices and tools I had been taught over my first weeks helped me to work quickly through projects, and the understanding of Drupal I was given helped me to readily answer questions. I ended up gaining a lot of experience working with Drupal and solving problems using the resources at hand. By presenting me with powerful tools for web development and a solid foundation with Drupal, Alister and Louie helped me to become both a more productive and a more self-reliant web developer.

Throughout my time with the IIC, people have always been extremely welcoming to me. I gave my first presentation to a group of project leaders about the enhancements I had made to the website. The purpose of the talk was to demonstrate the new content approval system I had created for the site that would allow everyone in the IIC to submit or edit content. I had met a few of the project leaders before at events such as the IIC Open House and some of the lunches, but I had not yet gotten to know most of them. Since it was also my first time showing off the progress I made since the start of the summer, I was nervous that I would encounter a new bug halfway through my talk. However, everything went smoothly, and afterward a handful of people introduced themselves and asked if I could take a look at redesigning their sites with Drupal. Also, Hanspeter Pfister, who runs the Connectome project and is a Harvard computer science professor, offered to be my advisor for the coming year. Less than a month into the summer, and I already had a queue of projects lined up and an advisor for my freshman year! Since then, I have continued working at the IIC, and having an advisor within my concentration has been extremely helpful.

As I finished my summer internship and began to work part-time over the school year, I continued to enjoy the lively atmosphere at the IIC. While finishing up my work on the website, I had begun to do work specifically for the Time Series Center. My goal for this project was to make an interactive linegraph viewer for browsing astronomical lightcurve data. With support from the National Science Foundation’s program on Research Experiences for Undergraduates, I have continued with this project over the school year.

Although I can do the majority of this work out of my dorm room, I still take every opportunity possible to go to the IIC. I always leave with some new idea or knowledge. Sometimes my visits to the IIC pertain specifically to my work for the Time Series Center, and so I discuss possible improvements for the graph and collect tips on how to best solve problems that I am struggling with. However, my conversations with the IIC staff are rarely limited to my current projects. Through my random discussions with Greg Rundlett alone, I ended up learning the ins and outs of Subversion code repositories and picking up tips and tricks for using Linux. When I became interested in programming...
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. A related effort, the SIOC (Semantically Linked Online Communities, sioc-project.org) project, provides semantics to a community but does not address the need for annotation with biological resources. There is an ongoing collaboration between SIOC and SWAN teams to align the ontologies.

The first instance of the SCF framework is being used to create an open access digital community for stem cell research—StemBook. Since its launch in September 2008 with 12 chapters, StemBook has grown to 39 chapters and has a current total of 89 commissioned chapters. StemBook is an editorial board-reviewed compilation of articles on stem cell biology. 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 the *Nature Network*

for the iPhone, Rahul Dave of the Time Series Center showed me some online resources that proved to be a much better starting point than the Apple documentation I had been plowing through up to that point. When I mentioned a difficulty I was having with a side project I was working on, I came out of a talk with Doug Alan with a new algorithm to try, which ended up solving my problem perfectly. For me, the IIC has not only acted as a source for gaining experience programming, but as a valuable resource of computer science knowledge and a major influence in developing my new interests.

Finally, my internship with the IIC helped broaden my understanding of disciplines beyond computer science. I gained knowledge on a wide range of sciences through my discussions with members of the IIC. My work with the Time Series Center alone has helped me learn many of the basic terms and theories used in astronomy. Although a knowledge of some astronomical principles was necessary for me to complete my work, I found myself learning about tangential subjects through talks with other people in the group. By attending presentations given by other interns within the Time Series Center, I grasped an understanding of some other fundamental topics in astronomy, such as the various types of stars and how the type can be inferred from a lightcurve.

Having come to Harvard already determined to concentrate on computer science, I had accepted that, apart from core courses, I would spend my time focusing on CS and math. My experience with the IIC has afforded me the opportunity to learn about other areas of science that I would otherwise have been unable to study while still maintaining a focus on computer science.

Tom Buckley has just completed his freshman year at Harvard. Tom Buckley's Harvard student experience started with an internship with the IIC Design Group. He has helped build web tools for the SCF, the IIC website and the Time Series Center (for which he created the graph interface below) with Harvard and NSF support.
IIC Projects

(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, 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.

Another SCF based project launched in May 2009 is PD Online Research, an online community of Parkinson’s Disease researchers. PD Online Research is a project of our group at MGH in collaboration with the Michael J. Fox Foundation (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.

Several other sites are under development, including a new web community for Neuropathic Pain research in collaboration with the Harvard NeuroDiscovery Center (Adrian Ivinson and Joe Martin), and a re-engineered version of the popular Alzheimer Disease research community Alzforum (www.alzforum.org). The sites developed on the SCF platform are interoperable with each other.

Publications and posters


Recognition


External relationships

Digital Enterprise Research Institute (DERI), National University of Ireland (John Breslin, Alexandre Passant) and Eli Lilly & Co (Susan Stephens) We are working with this group (a) to integrate the SWAN ontology with DERI’s ontology of blogs, wikis and discussion groups (b) to extend DERI’s semantic search capabilities to SWAN knowledgebases and SCF web communities, with a goal of specifically enabling cross-site semantic searching across Alzforum/SWAN, PD Online and StemBook as a model “community of communities”; (c) to further develop semantic web support within Drupal, generalize it and incorporate it into Drupal Core.

University of Rome “La Sapienza” We are co-developing a proposal to build an SCF-based web community for researchers in tinnitus.

Michael J. Fox Foundation (Michael Rogan) We are collaborating to build PD Online, an ambitious new web community for PD researchers based on SCF and SWAN; our Neuroscience 2008 poster on PD Online was selected by the Society for Neuroscience as an official “Hot Topic” poster for the conference.

Accelerated Cure Project (Art Mellor and Hollie Schmidt) We are developing a proposal to a major biotech foundation to develop a new web community based on SCF for multiple sclerosis researchers.

Nature Publishing We have agreed to work with the HSCI StemBook team, the editorial staff at Nature Stem Cell Reports, and the informatics group at Nature Publishing, to develop a common ontology of annotation terms for stem cell research, and to enable these to be used in annotating both StemBook and Nature Stem Cell Reports.

Harvard Office for Scholarly Communication, Elsevier Publications (Anita DeWaard), Digital Enterprise Research
and with other sites on the Semantic Web. In this new paradigm, there is a significant reduction in artificial barriers between research disciplines, and a much more dynamic and agile approach to information exchange. 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 (repositories of facts about things) and from “discourse-style” repositories such as SWAN (repositories of facts about assertions) as well as generating content via the actions of participants and editors. This content can be fed back into SWAN-style knowledgebases 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 are using it in developing a small ecosystem of communities in the neurodegeneration/neuro-repair specialist areas.

The SCF project is funded by private and government funding agencies for the PI, project manager, three developers and a postdoctoral fellow and will continue at the MIND Institute at Mass General Hospital in collaboration with SEAS, the Harvard Stem Cell Institute, the Center for Brain Science, and the Office for Scholarly Communication.

**Institute (DERI) / National University of Ireland** (Siegfried Handschuh), **USC Information Sciences Institute** (Ed Hovey), Centre for Language Technology, Department of Computing, Faculty of Science, **Macquarie University** (Robert Dale), **HMS Neurology** (Tim Clark). We are collaborating with this group to organize a conference, “The Future of Research Communication,” at Harvard in March 2010. SCF will be used to develop a “virtual conference environment” on the web, prior to and continuing after the physical conference, and showcase new technologies for research communications, the theme of the conference.

**Elsevier Labs** (Anita DeWaard), UCSD (Maryann Martone), BioMed Central (Ann Donnelly), Creative Commons (John Wilbanks). We are collaborating with this group to demonstrate new methods in service-oriented architecture for computationally-assisted semantic annotation of research communications, using SCF and the NIH Neuroscience Information Framework, for which Dr. Martone is the PI.

**California Institute of Technology**, Division of Biology (Wormbase / Textpresso Group, Sternberg Lab) We are collaborating with the Textpresso group to implement and support web services and annotator’s interfaces to their texitmining algorithms within SCF.

**Creative Commons / Science Commons** (Wilbanks, Ruttenberg et al.) We are working with the Science Commons project of Creative Commons to develop and integrate semantically characterized links to research tools information, held in a common repository, and applied as automatic annotations to biomedical content with SCF communities and open access publications.

**University of Southampton, UK** (David DeRoure) and University of Manchester, UK (Carole Goble) We have initiated a joint project with Profs. DeRoure and Goble to integrate the SWAN ontology used in SCF; and SCF semantically annotated content, with the myExperiment virtual research environment developed at Manchester & Southampton.

**Key collaborations**

**Harvard Stem Cell Institute (Brock Reeve and Lisa Girard)** StemBook, an online journal and web community of stem cell research, went “live” in September 2008 and will be indexed as a scientific periodical by the National Library of Medicine.

**HSCI Blood Disease Program** (David Scadden, George Daley et al. We are collaborating with the HSCI Blood Disease Program to develop an interoperable and reusable repository of genomic information with associated analysis methods as a component of SCF.

**Harvard Center for Brain Science** (Kenneth Blum, Executive Director); MassGeneral Institute for Neurodegenerative Disease (Janice Hayes-Cha, Executive Director) CBS and MGH-MIND are converting their website to run on SCF.

**Massachusetts Alzheimer Disease Research Center** (Brad Hyman, John Growdon, Brad Dickerson (HMS Neurology), Deborah Blacker (HMS Psychiatry), et al. We are working with MADRC to convert their website to run on SCF and to use SCF to provide a much larger set of content for MADRC researchers and the public about the Center.

**Joslin Clinic** (Chris Botka et al.) We are co-developing a proposal to a targeted donor to build an SCF-based web community for diabetes researchers.