# A. Overview and Structure of Proposal

OpenWetWare [http://openwetware.org/] is a collaborative web resource that provides biological researchers an online platform for storing, managing, and sharing primary and preliminary research data and know-how. By fostering the growth of online research communities, OpenWetWare has already enabled the organized capture of information and knowledge that is otherwise not stored electronically or disseminated. The work proposed here has two goals: first, to improve the organization of biological research knowledge information on OpenWetWare by developing tools that (a) support contextual tagging of information leading to improved knowledge quality and utility and (b) provide streamlined interfaces to existing databases; second, to develop a standard and open software distribution that can be used to jumpstart the formation of online research communities and thereby catalyze the development and promotion of standards of information and knowledge exchange across all fields of biological research. To accomplish these goals we are requesting resources to form a small team that will provide focused administrative, technical, and software development support to the OpenWetWare community; the requested resources will leverage the existing dedicated and energetic cadre of OpenWetWare volunteers. The proposal details the overall background and significance of the project [section B], preliminary results from the first 1.5 years of development and operation of OpenWetWare [section C], and details of the planned work, including a preliminary evaluation of models for long-term economic sustainability [section D].

# B. Background and Significance

### *B.I Information and Knowledge Loss in Biological Research*

The process of biological research generates and makes use of a wide variety of information, most of which is either inaccessible or unrecorded. For example, research papers and conference proceedings typically only summarize completed projects and results, while biological databases store and share heavily refined experimental data. Most of the remaining knowledge, and the process through which that knowledge is produced, is not recorded in any structured form or made available. As a second example, little or no information on failed research projects is disseminated [1, 2]. Moreover, most existing mechanisms for capturing, organizing, and sharing biological research information and knowledge are relatively expensive and slow, and are thus exclusive. For example, traditional publishing of research ideas, results, data, and knowledge is typically a month- to year-long process, often involving fees of over $1000 per manuscript. As a result, much biological research knowledge is consigned to individual and collective lore.

The limitations of lore as an approach to biological knowledge management, analysis, and dissemination are well documented [3-5]. For example, one recent study of academic-based genetics research found that 28% of researchers surveyed did not make available relevant DNA sequence information, 25% did not reveal all pertinent findings, 22% did not described detailed phenotypes, and 16% did not describe lab techniques [4]. The same study also showed that most researchers thought that the withholding of information and materials "slowed the rate of progress in their field of science" (73%) and had "adverse effects on their own research" (58%).

The reasons for poor information capture, management, and exchange in biological research are also well known. For example, 80% of those researchers that withheld knowledge responded that the effort to produce post-publication information or materials was too great [4]. Also, much of the relevant knowledge was thought to be inappropriate for inclusion in or along with a publication because of space restraints, lack of direct importance to the primary results, or because too many details might detract from the publication's main message. Once consigned to lore, biological knowledge becomes harder to share because there are neither strong incentives, nor an easy way to digitize, organize, and maintain such knowledge.

Two significant costs accrue due to the current approach to biological knowledge management, analysis, and dissemination. First, much biological knowledge is actually lost, resulting in a significant repetition of work across labs and over time. Second, most individual researchers are removed from and remain largely unaware of the challenges involved with knowledge management and analysis; thus, the individuals who are best suited to help specify the needs and requirements of biological knowledge tools, and who could collectively help to develop and implement distributed solutions to the challenges posed by biological knowledge, do not do so.

### *B.II Limits of Existing Biological Knowledge Management & Dissemination Technologies*

Biological knowledge is currently managed, analyzed, and disseminated via different mechanisms ranging from highly structured data type specific databases, to community-specific information portals, to individual laboratory and investigator websites. Highly-structured databases are currently used for DNA sequence and annotation information [6], protein structures [7], and large-scale experimental data sets [8]; the relatively fixed data structures underlying most large-scale resources prevents their rapid refinement and expansion; individual users and communities cannot readily tailor these resources to fit their specific needs. At the other extreme, community-specific resources such as WormBase [9, 10] and FlyBase [11], are heavily tailored to support the needs of a limited research group, but require large investment in centralized infrastructure [12]. The costs of developing large-scale databases and community-specific resources limits their use to relatively well-established, mature research areas. Emerging areas of research, new types of collaborations, and new approaches to knowledge management are excluded. Finally, individual laboratories often develop their own ad hoc knowledge solutions. Laboratory-specific solutions range from collections of physical media (e.g., index cards), to digital document archives, to custom-built databases or websites. However, these approaches are slow and usually do not support analysis of the knowledge itself or the reuse of tools that are developed for knowledge organization and analysis.

In addition to data type-, community-, and laboratory-specific resources, consortiums have emerged to develop standard methods for managing and sharing biological knowledge. For example, the Gene Ontology [13] and the Systems Biology Markup Language [14] provide standard languages for defining biological objects and relationships among these objects. Such efforts allow diverse groups to individually generate, share, and analyze biological knowledge without the need for a central database to store and access the underlying data. However, these projects require large dedicated efforts to define and refine their respective standards. While important, this process does not allow quick and easy ways to invent and test new standards. In addition, the standards themselves require dedicated tools in order to visualize, analyze, and share this information. Thus, consortium efforts share the same costs as community-specific resources, in that they require significant investments to support their maintenance and evolution.

Recently, tools have been popularized that enable massively collaborative knowledge management and dissemination [15]. For example, tens of thousands of users have collectively produced a useful encyclopedia of over one million articles called Wikipedia [16]. Wikipedia runs on an open-source software wiki distribution called MediaWiki. A wiki is a piece of software that allows many people to easily generate, edit, and link between online content simultaneously [17]. The in-line editing interface and simple syntax provided by most wiki software greatly reduces the effort required to contribute and edit knowledge, biological or otherwise. Furthermore, depending on the user management structure, wikis allow many individuals to collectively share and combine knowledge. As a second example, the "semantic web" effort is seeking to support automated information exchange via support for the inclusion of computer-processable meaning with already human-readable online knowledge . The semantic web can be used to provide a set of common standards that describe and name the relationships defined and described in text – for example, "this gene encodes this protein, which is active in this process." Built on the success of the web's hyperlinks, the semantic web gives individuals the capability to assign greater meaning to digital knowledge. Stated differently, the semantic web gives some of the power of structured databases to unstructured knowledge resources by allowing for the open and decentralized extendibility of the underlying data structures and relationships themselves.

### *B.III Impact of Proposed Work*

The continued growth and development of OWW should result in seven significant contributions to biological research. First, scientists will have access to more detailed and organized biological knowledge [section C.II.a]. For example, information not typically published in scientific literature, such as control experiments and negative results, can be easily disseminated and analyzed. Second, OWW will enable new opportunities for collaborations across institutional, geographical and socio-economic barriers [section C.II.b]. Underrepresented groups that face obstacles to sharing information or publicizing their work are able to do so more easily. Third, the availability of detailed authoring and version histories for every OWW entry will provide a quick, easy, and free mechanism through which individuals can contribute their knowledge, opening up new opportunities for evaluating scientific contribution, merit, and impact. Fourth, educational materials will be increasingly easy to develop, share, and reuse [section C.II.c]. Fifth, by exposing students to OWW via their courses, we will train a new cadre of researchers to systematically and digitally document and share not only experimental results but also the detailed context for those results. Sixth, OWW supports the meta-level integration and annotation of existing biological knowledge resources (e.g., the Protein Data Bank); thus, OWW end users will be able to provide context, experimental details, and framing knowledge for pre-existing structured databases, just by using OWW to support their day-to-day research. Finally, funding agencies, civil organizations, and the general public will have unprecedented access to the process of scientific discovery and working descriptions of current biological knowledge.

## C. Preliminary Results

### *C.I Overview*

Having recognized the issues and opportunities introduced above, the graduate students in my lab decided that they needed to directly address the challenges of biological knowledge management, analysis, and dissemination that arise within the process of biological research. Thus, in May 2005, students in the lab started OpenWetWare (OWW). OWW is a wiki-based resource that supports the digitization, storage, sharing, and collective editing and curation of biological knowledge. Because of local interest and in order to increase the quality of OWW's shared resources, we offered access to OWW to other labs at MIT, and soon thereafter to the entire biological and biological engineering research communities. Over the last year, OWW has grown

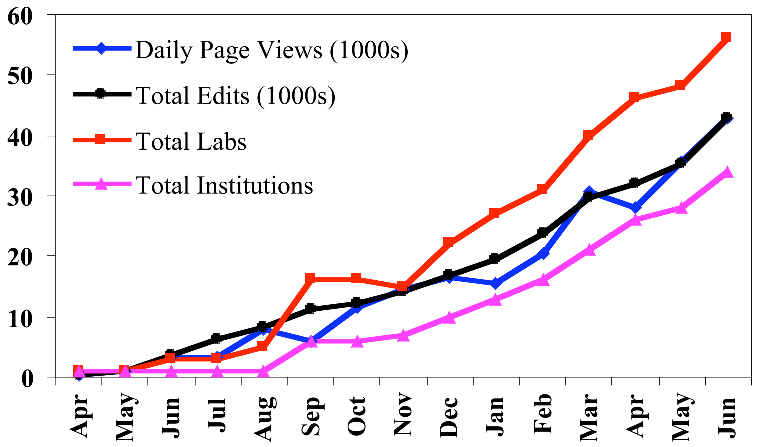

**Figure 1.** Growth of OWW from April 2005 through June 2006. The number of participating labs, institutions, and entry edits are cumulative. Page views are given as a daily average in thousands (e.g., in June 2006 OWW received an average of over 40,000 page views per day).

organically to support over 60 laboratories and 1000 users from around the world [Figure 1]. The rapid growth of OWW is a direct result of its usefulness as a knowledge management platform in supporting both the process and results of biological research. The ease with which users can generate, edit, and link information, the accountability of contributions made to the site, and the ability to communicate and collaborate between OWW members have made OWW a valued tool for knowledge management and dissemination. As further evidence of OWW's success, a dedicated group of volunteers has taken responsibility for ensuring that the infrastructure supporting OpenWetWare has grown along with the user base. Section C.II, describes how OWW has already been used to generate and curate types of biological knowledge not commonly found elsewhere. Section C.III describes the importance of communities on OWW and the technical infrastructure that has been built to support their growth.

### *C.II What is OpenWetWare?*

OpenWetWare (OWW) is a wiki-based resource that is focused on the capture, curation, and dissemination of the day-to-day working knowledge of biological researchers – knowledge that is otherwise lost in offline lab notebooks or shared only within small communities. OWW builds on the standard MediaWiki distribution via extensions that support the specific needs of biological research communities. Visitors to OWW, including the general public, are able to view all content. However, an authenticated user registration and login process is required in order to gain knowledge creation and editing privileges. OWW is structured to include areas that allow users to post content about themselves (User Pages), their laboratories and collaborations (Lab Pages). In addition, there are specialized areas that allow users to collaborate and improve general information resources (Shared Resources) and to comment and help improve OWW itself (Community Portal).

Three key software features and administrative decisions haven driven OWW's growth and distinguish it from the existing classes of biological knowledge management tools described above. First, editing content on OWW is very easy, and does not require knowledge of HTML. The underlying wiki software uses a simplified annotation that gives users an easy way to add knowledge, provide simple structure to the entry, and links to other entries. Second, all users can edit any entries on OWW, which allows for the decentralized editing and collaborative organization of biological knowledge. Third, because all knowledge on OWW is world-readable the site encourages knowledge reuse and a research culture that accelerates the sharing and collaboration of ideas and results.

### *C.III Biological Knowledge on OpenWetWare, Summary*

OWW users have generated many classes of biological knowledge that cannot be found elsewhere including: up-to-date individual and laboratory research ideas and projects; notes, advice, and expected results on hundreds of biological protocols; laboratory notebooks detailing project ongoing experiments; information on equipment operation, modification, calibration, and control experiments; aggregated informational resources on biological strains and genotype information; collaborative project discussions and data; community-specific information portals for particular fields; safety information and procedures; and so on. In addition, faculty at MIT and other institutions have successfully developed and taught courses via OWW. Below, we give a few illustrative examples to introduce the types of biological knowledge resources that have been developed and are available via OWW.

### *C.III.a - Protocol, Equipment, and Biological Knowledge Resources*

OWW users have developed protocols and other knowledge integral to biological research because it improves their ability to store, reuse, share, and improve this knowledge. The protocol collection on OWW already has hundreds of protocols covering different areas of research, from bacterial culture, to DNA manipulation and analysis, to PDMS microfluidic chip design, to protein structure modeling. These protocols can be quite detailed, as there are no restrictions on space. For example, one popular protocol posted by a user involves making better gels for electrophoretic protein separation [18]. The gel protocol contains background knowledge for how the protocol was developed (gleaning information from patent literature, which is linked directly), a summary of why these gels work better (better pH and buffering, better storage, ease of running), detailed protocols on how to run the gel, and pictures of gels run with the protocol for information on expected results. In addition, since all OWW entries are linked to the initial author and all subsequent editors, there are built in points of contact for further guidance. In comparison, existing online protocol resources are relatively static, non-collaborative resources that do not support easy cross linking with experimental planning and results [19, 20].

Biological knowledge on OWW is not limited to what are classically considered protocols. For example, proper usage, calibration and maintenance of equipment present biological researchers with its own knowledge management challenges. Unfortunately, knowledge on proper operation of laboratory equipment is typically limited to user and repair manuals of varying quality that are sometimes available from the manufacturer. In response, OWW users have started to develop equipment-specific knowledge entries. For example, the entry for a 96-well microplate reader contains knowledge on machine programming; data from control experiments on detection limits, linear range, lamp energy, and plate-to-plate variation; operational methods for responding to sample liquid evaporation and advice on optimal sample layout; scripts in Matlab and Excel for data analysis; and a service history for one lab’s machine, detailing major failures. OWW now contains dozens of similar equipment pages [21]. While equipment operation knowledge may seem trivial, capturing the detailed operation of the equipment used in biological research is necessary to support the broad development of accurate data models describing the results of laboratory experiments; such models, in turn, will be needed to support the extension of automated knowledge management, analysis, and dissemination tools to data types that are now outside the scope of the NSF Biological Database and Informatics program.

While OWW makes it easy for individuals to contribute knowledge, the unique strength of OWW becomes apparent when multiple researchers are able to collaborate to improve a single knowledge resource. Two current trends have emerged that demonstrate how users collaborate on OWW to refine biological knowledge. First, users aggregate data from different sources in order to provide a more detailed knowledge resource. For example, several labs posted protocols for DNA ligation using different methods [22]. Some members of those labs began a "meta-protocol" page describing the background and general procedure of DNA ligations, linking protocols from multiple laboratories and including a description of differences; other individuals later added tips, observations, and publications related to or based on the protocols. As a second example, another researcher created an entry detailing *Escherichia coli* genotypes [23]. Later another user-contributed explanations of the cryptic phenotype nomenclature allowing those outside the field to more easily understand the information on the page. The page now includes over 60 explanations of the nomenclature, information on over 40 commonly used *E. coli* strains, other information dealing with DNA methylation and other practical issues, links to related resources, and references to particular papers of interest.  Second, users are providing feedback of their experiences using other users' protocols. For example, a researcher posted a particularly detailed protocol on a method for quantifying proteins using a ß-galactoside assay [24]. Another researcher subsequently posted her general experiences with the protocol, sample data demonstrating the repeatability, and general levels of output to expect on a control experiment. Third, users are collaborating to aggregate biological information from disparate sources.

### *C.III.b – Communities*

OWW provides an easy to use and flexible method for developing online communities. OWW communities range from individual laboratories to multi-institutional groups working in the same field. The most common community on OWW is the individual research lab. For example, Pamela Silver's group at Harvard Medical School is an excellent example of a lab that has integrated OWW into their day-to-day research. They use OWW to share knowledge such as lab meeting schedules, protocols, research directions, and details on reagents [25]. The tools made available by OWW enabled the Silver lab to rapidly create an online community site for their lab and to populate it with novel information that otherwise would not have been disseminated.  A second type of community on OWW is devoted to fields of research. For example, synthetic biology is a new field emerging at the interface of science and engineering. Synthetic biologists have used OWW as their primary community knowledge portal; the portal itself is generated dynamically and automatically from the community-edited OWW entries [26]. Specific uses of OWW in the synthetic biology community have included the dissemination of news relevant to the entire community, discussion pages about experimental protocol standardization, discussions of new research projects and efforts, individual and group research results, links to community resources, conference and job announcements, links to individual labs on OWW in the field, and public discussions of societal issues facing the field.

### *C.III.c – Education*

The advantages of OWW that make it useful in research also make it a powerful platform for knowledge management in the classroom.  Already, there have been several initiatives for teaching classes through OWW.  Developing and maintaining course materials via OWW enabled collaboration on course development, facilitated student involvement and interaction, and allowed curriculum reuse.  For example, during the spring of 2006, MIT's Biological Engineering department taught a new undergraduate introductory laboratory class titled Laboratory Fundamentals of Biological Engineering (20.109) [27]. A team of four faculty, two instructors, and four teaching assistants taught the course. Together they developed the course content, which ranged from background materials on the particular course modules and experiments, detailed protocols, day-to-day laboratory instruction, and general information on the laboratory, safety, course policies, and presentations of earlier results. At the start of the course, each student was given an account on OWW. The students quickly began improving content on the site by identifying and correcting errors in course content [28], and uploaded experimental results based on the protocols [29]. Instructors were able to give feedback on those results and improve course material for future teachings [30]. Providing course content in a reusable form on OWW promotes sharing of educational ideas and materials within the community in a way that static course websites cannot. Moreover, the ability of students to provide peer-generated explanations for how protocols actually work provides a rich learning resource for subsequent students.

### *C.IV Technical Support and Community Organization*

OWW has relied upon community-driven leadership to identify and address the needs of active researchers. In C.IV.a, we describe how the organization of OWW leadership has evolved into a functioning steering committee. Next, we will discuss how the committee has handled and encouraged the growth of the overall community [section C.IV.b]. Finally, we will discuss the information and technical infrastructure that has been put in place in response to the changing needs of a rapidly growing community [section C.IV.c].

### *C.IV.a Establishing Community-Based Leadership*

Students in my lab initially managed all aspects of OWW organization and development. As OWW grew to include other individuals and labs, we established the OWW steering committee in January 2006. The steering committee is charged with guiding the future growth and development of OWW. Membership in the steering committee is on a volunteer basis. One member serves as an organizer, setting an agenda for a monthly meeting. Other members volunteer to spearhead particular projects. For instance, members have led small teams working on advertising, community development, software development, information management, coursework integration and other needs as they have arisen. The steering committee currently consists of 28 members from 12 institutions and is largely made up of graduate students. Decisions are made by consensus whenever possible or else by majority vote of members present (or teleconferenced) at the meeting. Since its inception, the steering committee has provided the overall vision, division of labor and community building for the OWW project.

In addition to the steering committee, there are also users who are very active editors and contributors to the site (termed power users). Power users identify problems and find solutions to ensure that OWW continues to meet community needs. They often suggest new initiatives to the steering committee and pioneer the addition of new types of knowledge entries on OWW. Furthermore, power users help new users become familiar with the site playing a role in user retention. As the OWW user base increases, the number of power users must scale accordingly.

### *C.IV.b Encouraging Growth of OWW Communities*

The researchers on OWW, led by the steering committee, have actively encouraged the growth of communities on OWW for several reasons. First, researchers often contribute knowledge in order to share within their own communities, and as a result of the knowledge being open, others benefit. Second, the recognition bestowed by community members on individual contributions provides motivation for researchers to contribute more knowledge. Third, communities that rely on OWW have a vested interest in making the site a better resource for their purposes.

One of the biggest concerns that new communities have when joining OWW is about vandalism and accountability, since every user on OWW is able to edit every entry. We have addressed these concerns on both technical and community levels. First, each new member of OWW is screened by a small group of steering committee members (aided by a custom user management tool) to ensure that only researchers that wish to add knowledge are given access and that each member has verified contact information. This also provides an opportunity to provide news users with a structured introduction to OWW. Second, each edit to OWW is tracked and stored on a history section associated with each entry. Vandalism, and for that matter mistakes, can be quickly reverted, and attributed to a particular user that had gone through the screening process. This not only provides strong accountability on each edit, but also provides a mechanism for giving credit for contributions by a researcher. Third, the steering committee established community guidelines asking users to refrain from making major changes to entries prefixed with a lab or group name. The effectiveness of these measures is illustrated in the fact that to date, OWW has not been subject to even a single identifiable case of vandalism.

The steering committee has led many other efforts to encourage community growth. The Getting Started pages provide new users with detailed tutorials on building personal and laboratory information sites, descriptions of what types of information laboratories often contribute, and explain how to become a power user [31]. The reorganization of the Main Page organized the large number of resources available on the site and helps attract new communities. The OWW Highlights were started by a steering committee member to call attention to special news and outstanding contributions to OWW in order to provide others with examples for use [32]. Finally, committee members and others communicate with new communities on the site to help with basic problems and inform them of useful features.

### C.IV.c Growth of Informational and Technical Infrastructure

The members of the OWW community have developed many software tools to enhance the user experience and usefulness of the OWW; the steering committee has taken steps to encourage the work of volunteers, such as hosting a parallel development site that allows users to develop and test new tools. Some of the particularly useful tools developed to date include the ‘Dewikify’, ‘Biblio’, ‘Filtered Changes’, and ‘Wiki Import’ extensions. The ‘Dewikify’ tool enables a page on OWW to be shown without the wiki frame surrounding the content enabling labs to use OWW both as an knowledge management tool and as their public, more aesthetic webpage. The citation manager, Biblio, enables the easy creation of citations and bibliographies within a wiki document [33]. The user simply provides a PubMed ID or ISBN number and a full citation with authors, publication title and reference, and links to PubMed is generated automatically. The ‘Filtered Changes’ tool allows users to filter recent changes to OWW by user, laboratory, and other useful criteria enabling a laboratory to easily monitor changes within areas of interest. Finally, the ‘Wiki Import’ tool allows a separate Mediawiki-based wiki to be merged with OWW providing a method to move all the contents of already-existing wikis onto OpenWetWare.

However, some tools are unlikely to be developed by volunteer efforts due to the complexity and effort required. For example, we contracted a paid developer to complete work on a new user management system with seed funding from the MIT/Microsoft iCampus program. In the future, moving some of the needed OWW software development to a small team of focused workers will allow development of important new tools, while also allowing the OWW steering committee and users to focus on other critical knowledge management tasks.

As OWW has grown, we have successfully scaled the computer hardware serving the site from a server in our laboratory to a faster server that is professionally managed off-site with regular backups. MIT's Computational and Systems Biology Initiative (CSBI) has graciously donated funds to allow us to remain on this faster machine [letters of support], however based on our current growth we will soon require increased hardware support [section D].

# D. Proposed Research

*D.I Overview*

The proposed work seeks to build upon the existing OWW community in order to achieve three specific aims: (1) establish a scalable organizational leadership for identifying and addressing community needs, (2) develop a series of critical tools for better knowledge generation, management, and dissemination, (3) implement and distribute new approaches to growing and strengthening online scientific communities and the knowledge they make available online.

*D.II Specific Aim 1: Community & Technical Infrastructure*

This grant will enable OWW to transition from a laboratory project to a self-sustaining community independent of my laboratory. The future organizational structure for site leadership is designed to facilitate this transition.  We also request support for technical infrastructure to meet estimated server and bandwidth demands.

*D.II.a OpenWetWare Leadership*

OWW leadership will consist of two components: the existing volunteer community-based steering committee and a new administrative/technical team (funded by this grant) tasked with supporting the committee. To date, the steering committee has shouldered administrative and technical burdens in addition to providing community leadership. This arrangement is not sustainable since the time and expertise necessary for OWW system maintenance and software development is outpacing what steering committee members can do on a volunteer basis. In the new arrangement, the steering committee will continue to identify and address community needs, and the administrative/technical team will support these efforts with specialized technical skills in software development and project management. The steering committee will continue to make decisions by consensus whenever possible and by majority vote if necessary. As the site matures, we will restrict the steering committee's size and the OWW community will select committee members by vote. Other community sites such as Wikipedia have successfully used such a system [34]. Table 1 provides a detailed description of the responsibilities of the project support staff, two developers, steering committee and myself in years 1-5.  Note that while the primary driving force behind OWW development is the steering committee, I will assume direct control over decision making if necessary.

|  | Year 1 | Year 2 | Year 3 | Year 4 | Year 5 |
|---|---|---|---|---|---|
| **Principal Investigator**<br>*Drew Endy* | - Hire Project support staff and Developer 1 |  | - Hire/promote full-time Project support staff | - Decide on long term financial sustainability plan |  |
|  | - Oversee developers' work<br>- Assess progress in meeting software dev. goals. |  |  |  |  |
|  | - Ultimately responsible for all expenditures<br>- Serve on Steering Committee |  |  |  |  |
| **Project support staff** (half-time, years 1-2; full-time, years 3-5) | - Aid Principal Investigator in overseeing developers' work and assessing their progress on meeting software development goals. |  | - Galvanize OpenWetWare communities at other institutions via seminars and tutorials.<br>- Oversee developers' work and assess their progress in meeting software development goals. |  |  |
|  | - Schedule, run, and record Steering Committee meetings.<br>- Conduct formal assessment of OpenWetWare's usage.<br>- Coordinate with TechSquare, Inc. on software and hardware purchases/updates/backups/maintenance. |  |  |  |  |
| **Developer 1** | - Develop extensions for automatic inclusion of biological data from existing structured databases based on reference ID (i.e. accession number, PDB code etc.)<br>- Develop tools for categorization of pages.<br>- Implement protocol submission wizards. |  | - Develop semantic web technologies for richer annotation of data in OpenWetWare.<br>- Other future projects TBD based on community feedback and steering committee recommendations. |  |  |
|  | - Attend and participate in Steering Committee meetings. |  |  |  |  |
| **Developer 2**<br>*Ilya Sytchev* | - Develop, document and release OpenWetWare MediaWiki software distribution. |  | - Incorporate new extensions from community into OpenWetWare MediaWiki software distribution for regular releases.<br>- Provide technical support for outside communities using distribution.<br>- Develop additional distributions (for instance, educational) as need arises. |  |  |
|  | - Attend and participate in Steering Committee meetings |  |  |  |  |
| **Steering Committee** | - Identify, prioritize and assign key areas of work needed for OpenWetWare site and community development.<br>- Prioritize and task software tool development with Developers.<br>- Serve as beta testers for new software tools.<br>- Bring concerns and recommendations of OpenWetWare users to Project support staff and Developers.<br>- Generate, curate and organize content on OpenWetWare.<br>- Make recommendations to PI on allocation of monetary resources. |  |  |  |  |

**Table 1.** Timeline and milestones for proposed research.

### *D.II.b Assessment*

We intend to evaluate the efficacy of the new tools developed for OpenWetWare. For example, are the new wizards increasing generation and standardization of content? Are people using the semantic web tools to annotate their knowledge entries? We will carry out project assessment in two ways. First, the assembled technical team will develop metrics to evaluate features based on simple scripts that automatically analyze entry edits and views. In addition, they will develop user surveys to be filled out by community members. Finally, the steering committee and power users will also provide important feedback on the utility of features as they have a strong grasp of the overall community and its needs. Based on all these inputs, the steering committee and technical team will evaluate and modify ineffective tools.

### *D.II.c Technical Infrastructure*

OWW has grown at a rate of ~20% per month for the last 8 months [Figure 1]. To support continued growth at this rate OWW will need investments in technical infrastructure. We will purchase new server equipment and hardware support in year 1 to meet our immediate needs, as well as in year 3 to account for projected server loads from the current ~20 GB per month to ~150 GB per month in year 5. We have an existing relationship with Tech Square, Inc. who has provided support for the servers donated to OWW by the Computational and Systems Biology Initiative at MIT. We will expand this relationship with explicit funding for new equipment, and a service contract to support hosting, management, and backups. OWW upgrades will be coordinated and vetted by the technical team described previously [section D.II.a].

### *D.II.d Long Term Sustainability*

The costs of maintaining OWW beyond this grant's funding period would primarily be in supporting the small team providing administrative and technical support. This level of continued support could come from a variety of sources. For example, if OWW is successful at continuing to become a centrally important knowledge tool across biological research, continued support from public, private or institutional donations is feasible. Alternatively, we could consider online ad revenue as an option for supporting the site. Based on what comparable sites charge, by year 5, a single banner ad on the OWW main page may yield anywhere from $150,000 to $1 million per year. These estimates depend on banner ad charges ($5-$25 per ad per thousand page views) and linear extrapolation of current traffic trends (6-9 million page views per month in year 5) [35-39]. We will formally evaluate all financial sustainability options and decide what funding model will best serve the OWW community at the beginning of year 4.

Finally, the funding requested here will help transition OWW into a full-fledged open source community. OWW distributions will be maintained on SourceForge and will build on the latest versions of the MediaWiki distribution, and thus will remain free. Even in the worst-case scenario where OWW can no longer be supported, the knowledge within the resource could be readily exported, both technically and legally, and the software tools developed by the community will remain freely available.

### *D.III Specific Aim 2: Improved Knowledge Generation and Management*

The technical team will spearhead the development of tools that are of immediate use to the OWW community, such as interfacing existing biological databases with OWW [section D.III.a], automating knowledge organization using categories [section D.III.b], providing opportunities to associate greater context and meaning with data through semantic web technologies [section D.III.c], and developing software tools for easing content generation [section D.III.d].

### *D.III.a New Types of Knowledge on OpenWetWare*

As mentioned earlier, researchers often develop new knowledge resources on OWW by aggregating knowledge from disparate resources. The ability to directly incorporate knowledge from large structured databases such as GenBank [6] and the Protein Data Bank (PDB) [7] would give researchers more powerful tools to pull together knowledge in novel combinations. Thus, we will develop tools to initially couple OWW to GenBank, a widely used tool in biological research. The user community has heavily requested this feature. Successful development of this tool (and subsequent evaluation of its usefulness) will guide development of interfaces with other databases (PDB, WormBase, etc.) as requested by the community in the future.

GenBank provides DNA sequences of organisms, vectors, genes, and other biological samples. The first phase of development will pull knowledge from GenBank, based on GenBank Accession Numbers. This will provide a means for users to add knowledge to a particular DNA sequence. For example, entering <genbank:seq>NC_001604</genbank:seq> will seed OWW with knowledge such as the source organism (Bacteriophage T7), sequence length (39,937 bp), and relevant references. Users can then easily incorporate this information into an OWW page that aggregates information about Bacteriophage T7, describes current research, or outlines a protocol or other knowledge resource. The next phase will incorporate methods of viewing sequence and features directly within OWW, as one currently can do in GenBank. This would allow users to focus on regions of interest in order to highlight information specific to the research being conducted. The most important feature of this tool will be maintaining the ease of use that allows OWW to be harnessed by non-technical users.

### D.III.b Automated Knowledge Organization

One of the major challenges facing OWW is efficiently organizing and storing knowledge generated on the resource [section C.II.a]. For example, relatively few researchers spend the extra effort required to link their protocols to the Shared Resources section. Hence, while the total knowledge content has grown quickly, Shared Resources pages have grown at a slower rate. To address this challenge, we will provide users with the ability to easily categorize their own contributions. This will allow users to label entries via defined categories that either already exist or that they define themselves. Categories will then be used to generate meta-pages that dynamically organize contributed knowledge. Categorization would provide easier mechanisms for organizing and finding shared knowledge resources on OpenWetWare which we expect, in turn, will foster to increased use and new contributions.

To start, one member of our development team will implement a simple scheme for categorizing OWW entries. A tab at the top of each entry will provide users the option to "categorize" the entry. The user will then pick or create a new category. Finally, the ability to produce custom dynamic entries has already been demonstrated with other tools such as Recent Changes filtering, and the developer will extend these tools to allow customized category pages.

### *D.III.c Providing Context and Meaning to Information*

While the organization of individual pages is important, providing context and meaning to information will provide a more powerful knowledge tool. Scientists should be able to flexibly categorize, search, and discover digitized knowledge. For example, a researcher should be able to easily assign short text phrases (tags) like “p53” and “Huntington’s disease” to a particular paper as well as crosslink the supporting data, electronic descriptions of the protocols used, analysis methodology, and the biological materials involved. However, the real advantage comes from the broad use of a community – when another researcher uses the same reagent and the same “tag”, the information automatically connects itself to the previous experiment. The impact of such automated connectivity is significant. Just as the open-source programming model allows the community to organically grow the software code base for OWW, the “tagging” model allows the community to organically organize and add value to the knowledge that is created day-to-day in the laboratory.

The implementations of such tagging technologies into OWW can be accomplished by using the tools developed at the World Wide Web Consortium (W3C) called the semantic web [section B.II]. The technologies behind semantic web are theoretically simple and are based upon three standards developed by the W3C. First, the Resource Description Framework (RDF) allows individuals to make statements relating two objects in the form subject, verb, subject (e.g., ‘Paper A’ *uses* ‘DNA Ligation Protocol B’). Each of the subjects and verbs are actually Uniform Resource Indicators (URIs). The RDF Vocabulary Description Language, S, provides hierarchies for concepts and relations (e.g., ‘DNA Ligation Protocol B *is a* Protocol’). Finally, the Web Ontology Language (OWL) allows more complex forms of types and relations as well as the ability to merge different ontologies by defining equivalence (e.g., If a 'Ligation' *is a* 'Protocol', and a 'DNA Ligation' *is a* 'Ligation Protocol', then 'DNA Ligation' *is a* 'Protocol'). Here, our implementation challenge will be to hide the technical details of the semantic web behind an easy to use interface.

We will begin by using the categorization of pages [section D.III.b] as an initial set of seed data to provide context to. For example, Page Foo on OpenWetWare categorized as a DNA Ligation Protocol will be converted into RDF statements by asserting 'Page Foo *is a* DNA Ligation Protocol.' Ontologies derived from the RDF statements can then be codified using RDF Schema and OWL (DNA Ligation Protocol *is a type of* Protocol). When the statements and relationships are defined using these standard formats, we can then harness many of the software tools already developed for the semantic web [40-42]. After this proof of concept, the next stage of work will require moving from tagging entire pages, to data and knowledge within pages. During this time frame we will also coordinate with the developers who have recently started developing prototypes of incorporating semantic web into MediaWiki [41], as well as other projects who intend to use these tools (EcoliHUB and NeuroCommons) [letters of support]. Again, the most formidable task we will face is combining these powerful tools in a simplified way so as to encourage and support their effective use.

OWW is uniquely positioned to take full advantage of technologies surrounding semantic web and apply them to biology. First, OWW is specifically designed to capture the typically undigitized knowledge in the laboratory. This prevents recapitulation of existing database knowledge and allows tighter integration among existing database resources. For example, a researcher would create a semantic link between a protocol on OWW and specific protein in the PDB by specifying the PDB ID in their protocol. The PDB could later collect this information automatically and include a link on the PDB page to the protocol. Through this mechanism, end users on OWW are providing context, experimental details, and framing knowledge for pre-existing structured databases, just by using OWW to support their day-to-day work. Second, the large community of active researchers on OWW using and supporting this technology could jumpstart the adoption and development of powerful semantic web tools in support of biological research and knowledge management. Third, the interconnections and categorizations on OWW already provide strong initial substrate for creating semantic relationships. Fourth, biology is a constantly changing science, and therefore OWW's flexibility to develop and easily make use of definitions and relations is very attractive. As consensus emerges around particular methods of tagging information, such methods can more quickly be incorporated into more formalized frameworks, such as the Gene Ontology. Fifth, the OWW community includes key members now introducing semantic web into the life sciences community, including John Wilbanks, the first staff member focused on life sciences at the W3C [letters of support]. Sixth, one of the developers tasked with the project, Ilya Sytchev, has significant experience with both Semantic Web and wiki technology via his work on the MIT Registry of Standard Biology Parts [43]. Finally, we will coordinate with other large wiki-based projects interested in incorporating semantic into wikis, including EcoliHUB and NeuroCommons so that the tools we develop are generally useful to other groups through the OWW distribution [section D.IV; letters of support].

### *D.III.d Increasing Ease of Use*

Our initial efforts at improving ease of knowledge contribution have been successful [section C.III]. To further improve the ability of users to contribute and manage knowledge on OWW we will develop a framework that supports “wizards.” Wizards will provide generic templates for knowledge entries on OWW such as user pages, laboratory pages, protocols pages, equipment pages, and so on. We expect that wizards will not only increase ease of contributing knowledge, but also promote more structure to the knowledge as it is being produced, for example by increasing the use of tagging and organization technologies on OWW. For example, incorporation of some of the tools for organization [section D.III.b] and information tagging [section D.III.c] into the wizards will improve the use of these technologies and will accustom researchers to their benefits. As a result, the wizards will also serve the secondary purpose of enabling better organization and standardization of OWW-based knowledge.

To start, one of the developers will develop a protocol submission wizard and a detailed guide to OWW wizard development. The wizard will ask users for information about a protocol such as the title, category, tags, list of labs or individuals that use it, and list of materials that are involved. Using this information the wizard will generate a page for the protocol that follows a standard format, and automatically generate links from the listed labs and user page's protocols sections. Additionally, the wizard will place the protocol in the appropriate common protocols area based on its category and tags. The guide to the wizard development process will enable motivated users to build similar wizards for information relevant to their own communities.

### *D.IV Specific Aim 3: A Standardized Wiki Distribution for Biological Research*

OWW is committed to remaining a resource for open dissemination of knowledge related to biological research. There are many individual efforts to use wiki software for either storing and organizing private laboratory information or facilitating larger collaborative projects aimed at community work with specific goals such as genomic annotation [44]. In each case, investigators are building project-specific software. In order to allow OWW to serve as meta-level resource for exploring and implementing the integration of disparate types of biological knowledge, we will develop, distribute, and support a generic software distribution that includes the customized MediaWiki software and tools described in the previous sections.

We expect that an OWW organized software distribution will solve a number of problems. First, it will allow individual investigators to easily establish private wiki-based knowledge management systems. This benefits individual investigators and small communities by providing unfettered access to the many tools that make OWW useful, as well as reduces the effort for producing and maintaining such sites. Open efforts, such as OWW, benefit in return because investigators are putting their knowledge into an interoperable digital form that will make it much easier to make public when appropriate. The OWW distribution will contain a simple method to move (i.e., publish) knowledge from a private wiki to other wikis such as OWW. There are already several investigators, including our laboratory, eager to use the distribution to support their private work [letters of support].

Second, the OWW software distribution will enable several efforts by larger communities to develop wikis to generate and curate specific knowledge. These larger projects have their own needs specific to each community, and thus have embarked on using and developing their own wiki infrastructures. However, coordination between these groups would help ensure interoperability, reduce repeated efforts at tool development and customization, and allow strategic division of labor for specific goals. To begin, three new projects, EcoliHub, Wikiomics, and the NeuroCommons have agreed to coordinate tool development and distributions to ensure interoperability between these large projects in the future [letters of support].

The commitment to an open-source standards-based approach (e.g., MediaWiki and W3C's semantic web) between various biological knowledge projects is important because the ability to flexibly integrate new technologies is essential to any software distribution – imagine a web browser that cannot run media players, for example. While it is impossible to predict or program for disruptive technological advances, OWW's standards-based approach represents the best technical methodology to react to change. Furthermore, the open source methodology by which individual users can adapt the system to react to opportunity and re-contribute code means that the OWW distribution can grow and adapt without a costly organization driving requirements.

OWW is in the best position to lead the distribution effort because of the tools and extensions already developed specific to biologists, the experience of hosting a large scientific community, and the insight from that community to lead development of new useful tools. Ilya Sytchev, the programmer who has thus far led the technical upkeep and maintenance of OWW, will spearhead the project as one of the two developers on OWW. To begin, Sytchev will develop two distributions, one geared towards individual laboratories with privacy concerns, and another for open sites similar to OWW. These distributions will contain all the extensions, tutorials, help pages, etc. that make OWW easier to use and more powerful than the standard MediaWiki distribution. In addition, he will work on extensions that make it easy for these distributions to share knowledge amongst each other by means of a new "publish" tab. This will allow simple publishing of a private protocol page on an individual's wiki to OWW. Also, throughout the project, Ilya will act as a liaison with other projects to incorporate and vet tools developed elsewhere into OWW. Most importantly, Sytchev will ensure future compatibility of these extensions by testing these tools as new versions of the core MediaWiki software are released.

*D.V OpenWetWare & Education:* OWW is also a useful tool for developing and hosting educational content. Both courses under development and laboratory courses benefit from the collaborative and flexible editing of pages that OWW fosters. Three major barriers to the use of OWW for educational purposes exist. First, instructors don't know how to construct course pages on OWW. Second, it is difficult to move existing course materials to OWW. Third, students are unfamiliar with wikis. These barriers can be addressed in a multi-pronged approach very similar to those outlined previously. First, a course addition wizard, similar to the proposed protocol submission wizard, will be developed to streamline the entry of courses onto OWW. Software tools that facilitate conversion of word processing and presentation documents to wiki format will ease import of existing course materials. Specialized templates and tutorials, similar to those that already exist for general OpenWetWare users, will be developed to help in teaching students how to use the wiki for their own class work. If the demand arises, a custom OWW distribution for education might become appropriate. Given that OWW has only been used in a few courses to date, albeit quite successfully, our research plan in the context of education is necessarily exploratory.

*D.VI Statement on Intellectual Property:* All software developed under this grant will extend and integrate with the existing MediaWiki software distribution and therefore must fall under the terms of the GNU General Public License, a free software license. Individual OWW users already authorize licensing of the knowledge they contribute to OWW under terms of the Creative Commons BY-SA license, which allows automatic attributed re-use and extension of the contributed knowledge so long as any resulting works are also shared. By allowing both OWW tools and content to be free for reuse and sharing, OWW will be able to support the widespread development of similar biological research knowledge resources.