# Supporting Open Collaboration in Science Through Explicit and Linked Semantic Description of Processes

Yolanda Gil[1(✉)], Felix Michel[1], Varun Ratnakar[1], Jordan Read[2], Matheus Hauder[3], Christopher Duffy[4], Paul Hanson[5], and Hilary Dugan[5]

[1] Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292, USA
{gil, felixm, varunr}@isi.edu
[2] Center for Integrated Data Analytics, U.S. Geological Survey, Middleton, WI 53562, USA
jread@usgs.gov
[3] Software Engineering for Business Information Systems, Technical University Munich, Munich 85748, Germany
hauder@in.tum.de
[4] Civil and Environmental Engineering, Penn State University, University Park, PA 16801, USA
cxdll@psu.edu
[5] Center for Limnology, University of Wisconsin, Madison, WI 53706, USA
{pchanson, hdugan}@wisc.edu

**Abstract.** The Web was originally developed to support collaboration in science. Although scientists benefit from many forms of collaboration on the Web (e.g., blogs, wikis, forums, code sharing, etc.), most collaborative projects are coordinated over email, phone calls, and in-person meetings. Our goal is to develop a collaborative infrastructure for scientists to work on complex science questions that require multi-disciplinary contributions to gather and analyze data, that cannot occur without significant coordination to synthesize findings, and that grow organically to accommodate new contributors as needed as the work evolves over time. Our approach is to develop an organic data science framework based on a task-centered organization of the collaboration, includes principles from social sciences for successful on-line communities, and exposes an open science process. Our approach is implemented as an extension of a semantic wiki platform, and captures formal representations of task decomposition structures, relations between tasks and users, and other properties of tasks, data, and other relevant science objects. All these entities are captured through the semantic wiki user interface, represented as semantic web objects, and exported as linked data.

**Keywords:** Semantic MediaWiki · Open data science · Organic Data Science


F. Gandon et al. (Eds.): ESWC 2015, LNCS 9088, pp. 591–605, 2015.
DOI: 10.1007/978-3-319-18818-8_36

## 1 Introduction

The Web was originally developed to support scientific collaboration. Today, scientific collaboration over the Web takes many forms, including blogs, wikis, forums, code repositories, etc. These collaboration frameworks, like the Web, are used beyond science and are often originally developed outside of a science context.

We are interested in supporting scientific collaborations where joint work occurs on a concrete problem of interest, with many participants, and over a long period of time. Although the Web may be used to share information, there is no explicit support for the shared tasks involved. These tasks are discussed through email, phone calls, and occasional face-to-face meetings. We focus on scientific collaborations that revolve around complex science questions that require:

- *multi-disciplinary contributions*, so that the participants belong to different communities with diverse practices and approaches
- *significant coordination*, where ideas, models, software and data need to be discussed and integrated to address the shared science goals
- *unanticipated participants*, so that the collaboration needs to grow over time and include new contributors that may bring in new skills, or data

Such scientific collaborations do occur but are not very common. Unfortunately, they take a significant amount of effort to pull together and to sustain for the usually long period of time required to solve the science questions. Our goal is to develop a collaborative software platform that supports such scientific collaborations, and ultimately make them significantly more efficient and commonplace. Some scientific collaborations revolve around sharing instruments (e.g., the Large Hadron Collider), others focus on a shared database (e.g., the Sloan Sky Digital Survey), and others form around a shared software base (e.g., SciPy). In contrast, our focus is on collaborations where participants jointly pursue a shared scientific question.

We are developing a new approach to on-line collaboration that we call *Organic Data Science*. Our approach enables users to create tasks, exposes how they are being addressed, and facilitates other users to join in solving any task.

Our Organic Data Science framework is implemented as an extension of a semantic wiki, in particular the Semantic MediaWiki platform [1]. Users can add properties to tasks as needed, and can describe any entity of interest to the collaboration (datasets, software, papers, etc.) using semantic properties of the wiki. Semantic wikis provide an easy-to-use interface where users can define structured properties, which are then represented in RDF. The framework is still under development, and it evolves to accommodate user feedback and to incorporate new collaboration features.

There is a wide range of approaches that have been explored for collaboration, although they have not had much adoption in science practice [2]. There is also a significant body of work on studying on-line communities [3], notably on Wikipedia. Our work builds on the social design principles uncovered by this research.

The main contributions of this work are: (1) the design of the framework so it can capture structured information about scientific tasks and associated entities, (2) the implementation of the framework as an extension of a semantic wiki platform, and (3) the integration of the framework with other systems through the use of linked data.

This paper begins with an overview of the framework and the kinds of information captured to make the science process open. We then discuss the overall architecture and implementation of the system. After an overview of related work, we present some preliminary data on the use of the framework, and conclusions and future work.

## 2 The Organic Data Science Framework

Our approach is to expose science processes declaratively to support the formation of ad hoc groups to work on tasks of interest, to enable anyone to contribute to tasks that match their interests, and to advertise ongoing work to potential newcomers. Science processes describe the what, who, when, and how of the activities pursued by the collaboration. This section describes the Organic Data Science framework, focusing on how semantic representations are used. We use examples from an ongoing collaboration that is using this framework to study the age of water in an ecosystem,[1] but have anonymized the examples by using fictitious names.

The framework incorporates principles from studies of successful on-line communities, which we describe elsewhere [4].

### 2.1 Representing Tasks

Every task has its own page, and therefore a unique URL, which gives users a way to refer to the task from any other pages in the site as well as outside of it. Subtasks can be created that will be linked to the parent task, resulting in a hierarchical task structure. Task pages follow a pre-defined structure that is automatically presented to the user when a new task is created.

Figure 1 illustrates the representation of a task. On the left, the task is highlighted in the context of all its parent tasks as well as other top-level tasks. On the right, the subtasks are shown at the top. The bottom right shows metadata properties of tasks. As in any wiki page, text can be included to describe the task. Following the text, there is a space where users can define additional structured properties. Each task has an icon to the left that indicates progress on the task.

**Task Metadata.** Task metadata are major semantic properties of the task. We created a tabular interface to enter semantic properties. All task metadata is stored in the wiki as semantic properties of the task page.

We distinguish between *pre-defined metadata* and *dynamically-defined metadata*. *Pre-defined metadata* are properties of tasks that the system will use to assist users to manage tasks. *Dynamically-defined metadata* allow users to create new properties on the fly that help group tasks with domain-specific features, for example tasks that are related to calibration of models or outreach tasks.

Pre-defined metadata can be *required* or *optional*. Required metadata includes the start date, target date, task owner, task type (high, medium, and low level), and a

[1] http://www.organicdatascience.org/ageofwater/.

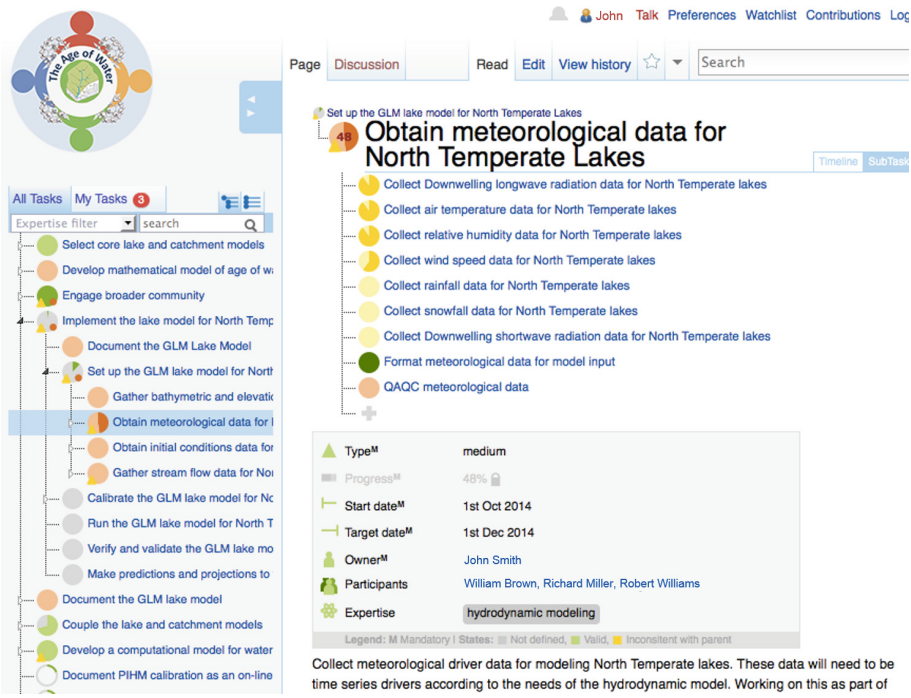
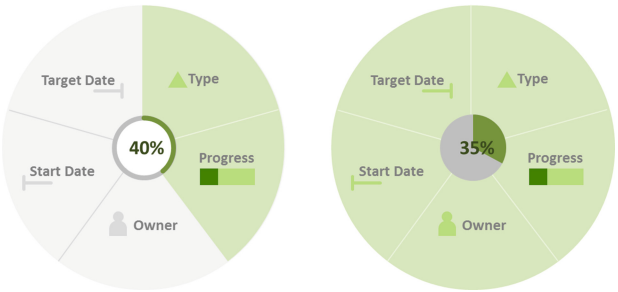



**Fig. 1.** Organic Data Science: describing tasks.

user-provided estimate of the progress to date. Tasks whose required metadata is incomplete have special status in the system and are highlighted differently in the interface to alert users of their missing metadata. Optional task metadata includes the task participants and the task expertise indicating the kind of background or knowledge required to participate in the task.

An important aspect of the framework is tracking the contributions of each user. This allows the system to show who can be credited for the content of each page.




(a) Task Metadata Progress (b) Task Content Progress

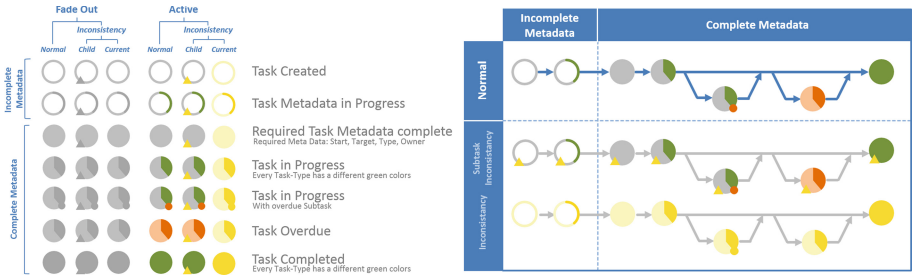
**Fig. 2.** Conceptual task state estimation.

**Task Status.** The system uses metadata properties to estimate the progress and status of tasks. Tasks that have a type indicated as high-level are assumed to have a high degree of abstraction and high uncertainty in the estimation of the task completion, such as the major tasks at the project level. Medium-level tasks are those that have a medium uncertainty in estimation of the task completion, such as activities within the project that are decomposed into several subtasks. Low-level tasks are those that have a low uncertainty in estimation of the task completion, such as small well-defined tasks that can be accomplished in a short time period.

The user selects the task type, which is indicated in the interface with different tones of green in the task icon. High-level task are colored in lighter green and lower-level tasks in darker green. The progress to date for low-level tasks is provided manually by their owners or participants, since the tasks have small duration. The progress of high-level and medium-level tasks is calculated by the system.

The progress of a medium-level task is calculated as an average of the progress of its subtasks. For high-level tasks, we assume a linear progress based on the start and target date in relation to today's date. This is because we assume that high-level tasks may have subtasks that have not been specified yet. To provide simple user feedback, metadata properties are shown in different colors to indicate their state: metadata properties that are not yet specified are shown in gray, valid properties in green, and properties that are inconsistent with properties of the parent task in yellow.

Figure 2 illustrates how the system uses the task metadata to generate the task state. The left side of the figure shows an example of a task whose required metadata is incomplete, where the Task State shows the percentage of required metadata that has been provided by users inside of a ring that shows that percentage in green. The side right of the figure shows an example of a task where users have provided all required metadata. Their status is represented by a pie chart showing the progress metadata property value in green. Different shades of green are used to express the task type, with lighter green indicating higher-level tasks.




(a) Possible task states.

(b) Sample transitions for task states.

**Fig. 3.** Task states and sample task state transitions.

Figure 3(a) illustrates all possible task state icons. The left columns show the task state for tasks which are faded out in the interface (shown just to provide context but do not match a search filter). Overdue tasks are indicated with an orange pie chart. A small orange point indicates that at least one subtask is overdue. This helps users identify overdue subtasks. Yellow icons indicate inconsistent tasks, which may be caused by

move actions, for example if their start date is before the start date of a parent task. The yellow triangles indicate an inconsistent subtask. Note that yellow and orange colors were also used to indicate overdue and inconsistent tasks.

Figure 3(b) illustrates some sample transitions for task states. For example, the first line shows a typical task that has no metadata when it is created, then required metadata is added but no work has been done in the actual task, and then progress in the task grows until completion although in some cases a subtask or the task itself can be behind schedule. The task state is shown in three different sizes depending on the location in the interface. Large size icons include the progress as a percentage, and are used for the currently opened task as well as in the user pages.

**Task Cloning.** We have found that often times the same task is done by several people with their own data. To support this, we have created a task cloning facility that takes a task tree and create personalized versions for a set of users. An example is shown in Fig. 4. The group held a workshop that had more than 50 participants, with the goal that each should be able to run a particular hydrological model with their own data. Several general tasks were created which documented what needed to be done in terms of installing software and prepare the data. The system then created personalized versions of those tasks for each workshop participant. This capability enabled the workshop organizers to track where each person was in the process. Each participant could annotate in their own page the particular problems that they were running into.


**Fig. 4.** Cloning tasks for several users to track their individual progress.

## 2.2 Representing Users

The system automatically creates a page for each user with an account on the wiki.

Figure 5 shows an example of a user page (broken into two pieces to fit the space). The system shows in that page the tasks that the user is owner or participant in, and organizes them according to whether the task is ongoing, upcoming, or completed based on the start and end dates. To do this, the semantic properties of the task are used. The system also retrieves all the expertise involved in the tasks that the user is contributing to, and shows it above the tasks.

The system also shows the most recent contributions made by that user to the different pages of the wiki (top right of Fig. 5). This is important to highlight the areas of the collaborative work that each user is working on. In addition, users can see their work recognized. The system also displays a scoreboard of credits in the front page of the wiki.

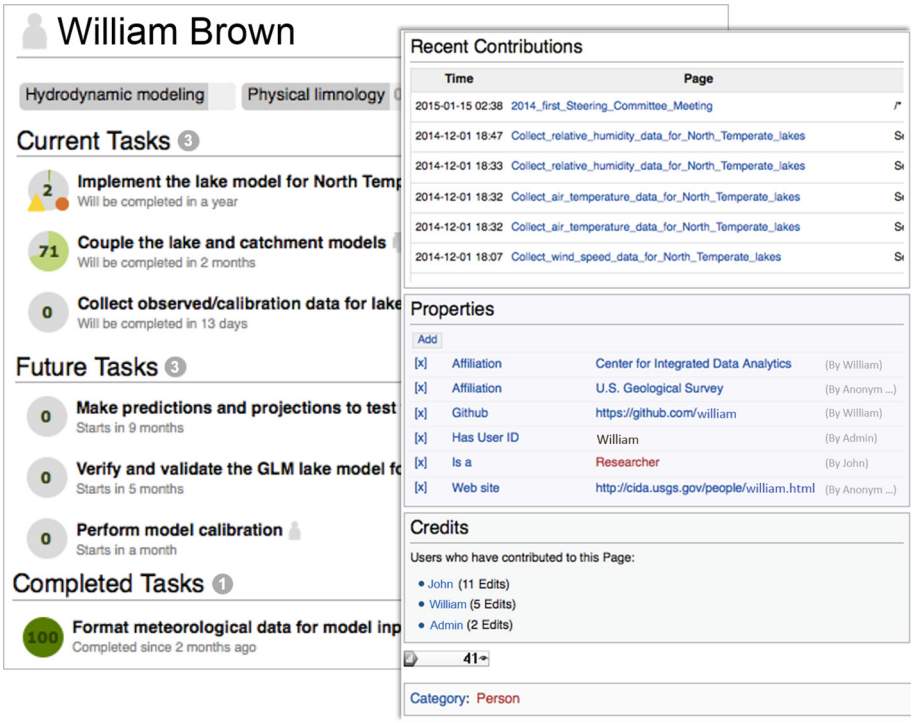



**Fig. 5.** Organic Data Science: describing users.

User pages can also have metadata properties. This is shown on the bottom right of Fig. 5. Properties can be added by other users, as is the case here. Credits are then shown to acknowledge all users that have edited the user page.

### 2.3 Representing Data, Software, Workflows, and Other Entities of Interest

Like tasks and people, any other entities of interest in the collaboration can be created to have their own page and associated URL. The most common entities are data, software components, and workflows, and we have created a pre-defined structure that is automatically presented to the user when a new entity is created.

Data is an important entity in a scientific collaboration. Figure 6 shows an example of a dataset description on the left. Datasets can have a type, in this case it is sensor data, and can have metadata properties. Users can add any metadata properties that suit their purposes in using the data. There are two major types of data. *User-described data* is stored in existing repositories external to the wiki. Users then just add a pointer (URL) to the dataset, and simply describe its metadata properties. *User-provided data* is uploaded to the wiki by users, and also described with metadata properties. This distinction enables seamless integration with external data sources.

In some cases, users will want to have default properties for some types of data. We have extended the framework to support this. An ontology of data types and default properties is used to create a customized property entry table, and users can always add additional properties separately as needed.

We are in the process of creating APIs to exchange information with data sources that would like to include the RDF properties captured with the Organic Data Science framework.

Software is another type of entity that are important in scientific collaborations. Figure 6 shows an example on the right-hand side. Software components have pre-defined metadata such as the inputs, parameters, and outputs. Users can add other metadata properties, such as the authors of the software, the language of its implementation, and pointers to the software repository.

Workflows are also important to capture the data analytics aspects of the work. Figure 6 shows an example of a workflow on the bottom. In this case, we show a reusable workflow template with links to the software components for each step. Workflow templates are also linked to their executions. Each workflow execution points to datasets (inputs, intermediate, and outputs).

We use a separate workflow system to generate workflows, then import them to show them in the Organic Data Science framework. We use the WINGS workflow system, which captures semantic properties of the data and workflows. WINGS exports workflow templates as linked data, as well as workflow executions using the W3C PROV standard. The workflows are then imported into the wiki. The process is described in detail in [5].

Other entities of interest can also be described in the wiki. For example, if the sensor data was collected for a particular location with a specific sensor, the location and the sensor can be described in detail through semantic properties.

### 2.4 Queries

All the semantic properties are stored in the wiki framework as RDF assertions. Semantic properties are queried in two important ways.

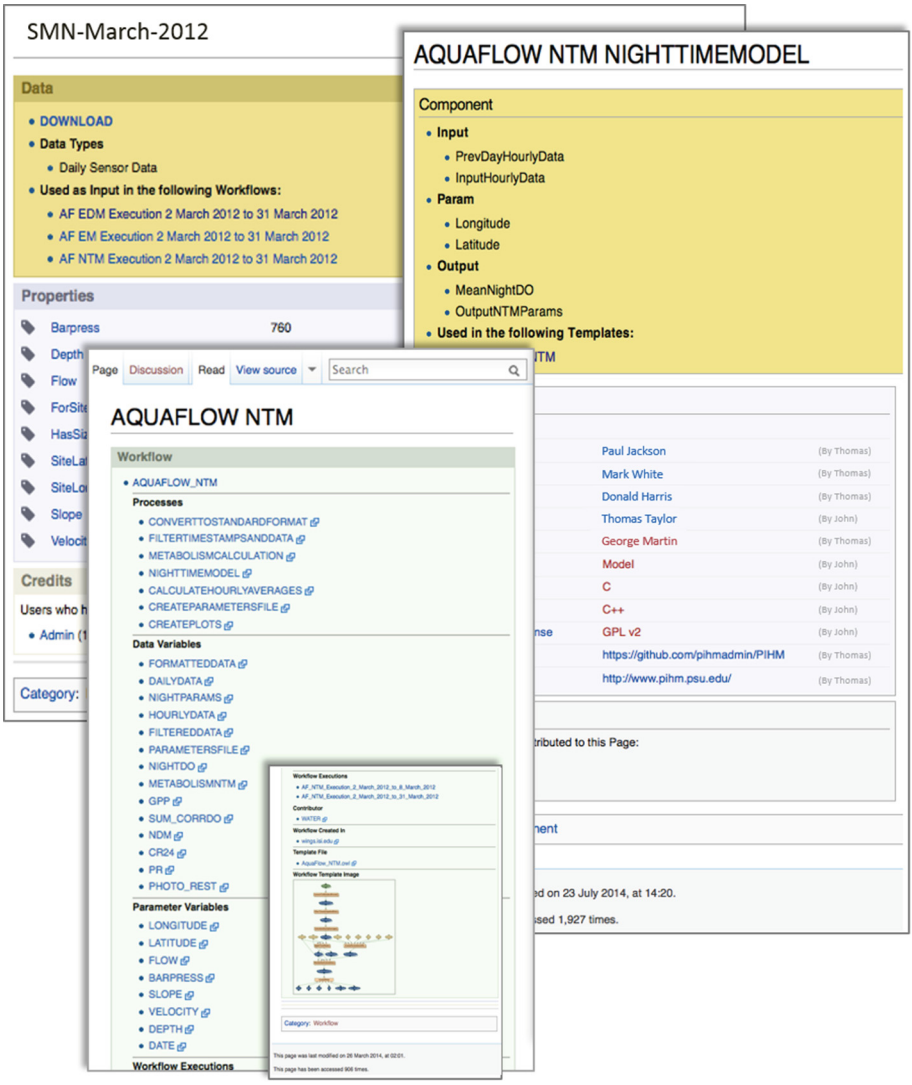



**Fig. 6.** Organic Data Science: describing datasets, software components, and workflows.

Semantic properties are queried by the system to assist users in managing tasks. We described in Sect. 2.1 how task properties are used to generate the status icons of tasks. They are also used by the system to generate much of the content of the user pages, as we described in Sect. 2.2.

Semantic properties are also used to generate wiki page content. Semantic MediaWiki offers a query language that can be embedded in a wiki page to dynamically generate content.

Figure 7 illustrates how the metadata properties of the task are used in queries. In this case, a dynamically-defined property "participant-of" was added to indicate the

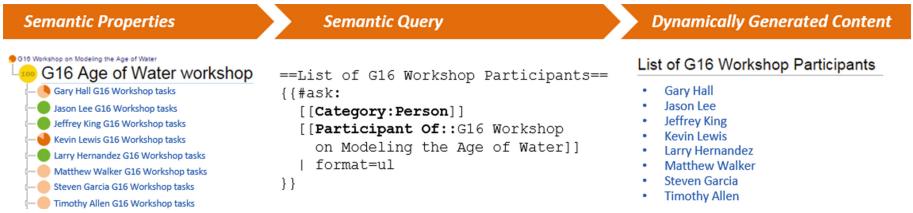



**Fig. 7.** Automatically content generation with semantic queries.

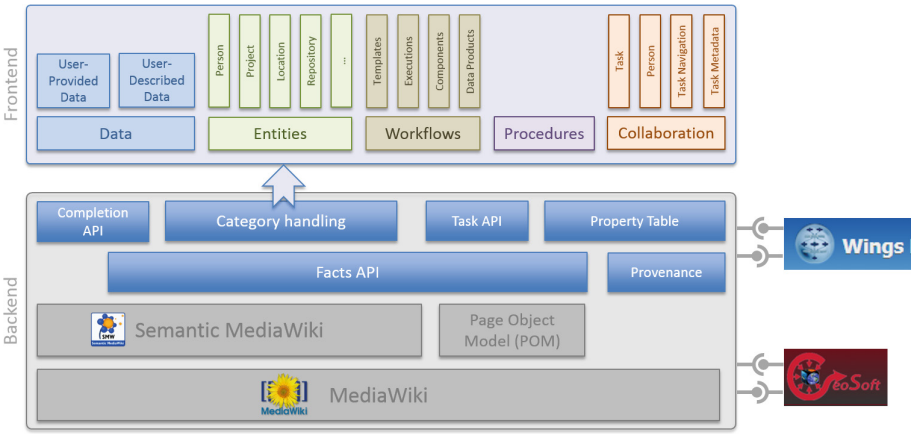



**Fig. 8.** Architecture of the Organic Data Science framework.

participation of people in a workshop. On the left we show the tasks that were involved in participating in that particular workshop. In the middle, we show the query in Semantic MediaWiki to extract all the participants. On the right is a page that is dynamically generated based on the users that participated in the subtasks created for the workshop.

## 3 Architecture

This section describes the architecture of the Organic Data Science Framework. A high-level overview of the architecture is shown in Fig. 8. The Organic Data Science Framework is implemented as a set of extensions of the Semantic MediaWiki and MediaWiki platforms. We also use the Page Object Model (POM) extension of MediaWiki,[2] which supports the manipulation of the content of the wiki pages. These three existing components, which provide underlying infrastructure, are shown in dark grey at the bottom of the figure. The rest of the components in the figure are the extensions that comprise the Organic Data Science Framework.

[2] http://www.mediawiki.org/wiki/Extension:Page_Object_Model.

We developed an extension to assert and retrieve assertions in the wiki, which is the Facts API. This enables easy access to the semantic properties regardless of how specific properties are handled in Semantic MediaWiki.

The Provenance extension handles attribution for each assertion in the system. Each semantic property is annotated according to the user that asserted it. This provenance information can be queried to generate the credit shown in the different pages.

The Completion API extension enables the system to offer users completions of the properties as they are typing, based on the properties that already exist in the system. This encourages users to adopt properties that others have already created, fostering agreement and normalization of property names. The Task API extension is customized to handle information about tasks. It manages the task-subtask tree, generates the status icons, and tracks task deadlines to generate user alerts.

Finally, the Category Handling extension manages the generation of different pages that are displayed to the user, depending on the category of the page. We described in Sect. 2 different categories of entities, such as tasks, users, data, etc. We have developed other categories at the request of users that are not discussed above, including procedures and data repositories. The representation of a person can be different, for example to distinguish someone who is part of the collaboration and should have a page as described in Sect. 2.2 from a person who has developed some software of interest but is not part of the collaboration.

The Organic Data Science framework can interact with external systems through the use of Semantic Web representations. We discussed above the integration with the WINGS workflow system.[3] Other external systems that we plan to integrate into the Organic Data Science framework include data repositories, software repositories, collaboration networks, and publication repositories.

The Organic Data Science Framework software is open source and is released on GitHub under an Apache 2.0 license.[4]

The Organic Data Science Framework can be set up for different communities. If communities choose to do so, they can make decisions to split the site into separate sites. Each site can point to others as URIs, enabling a looser form of collaboration. We have set up a special site for training new users. Each user is given a set of personalized training tasks, generated with the Task Cloning facility described in Sect. 2.1. Users are trained first to contribute to existing tasks, which is very simple training and takes 20–30 min. They are then trained to create new tasks and manage them, which is more advanced and requires another 20–30 min.

## 4 Use of the Organic Data Science Framework

The major use of our framework is by a community of hydrologists and limnologists that are studying the age of water in an ecosystem while collaborating with us to develop the Organic Data Science framework.

[3] http://www.wings-workflows.org/.

[4] https://github.com/IKCAP/organicdatascience.

Figure 9 illustrates the evolution of the collaboration graphs generated from the task metadata properties that link tasks and users. Each user is a node in the graph, with the links indicating whether two users have a task in common where they are owner or participant. The thickness of the link indicates how many tasks two users have in common. The graph on the left illustrates that many users collaborate with several others in different sets of tasks. It also shows that two different sub communities were being formed in practice (top and bottom areas of the graph), and the group agreed to split the work into two separate sites whose collaboration graphs are shown on the right of the figure.

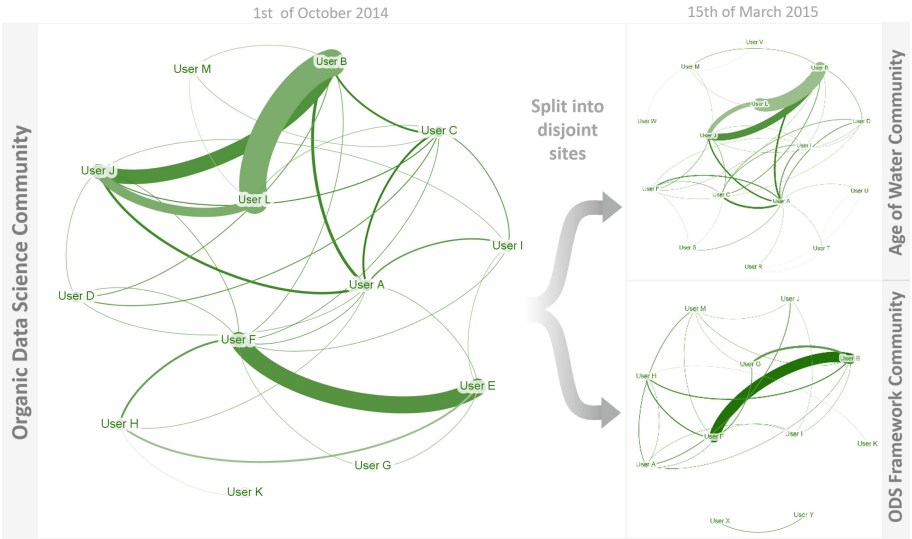



**Fig. 9.** Evolution of the Organic Data Science collaboration graph.

Several other science collaborations are starting to use the wiki:

- A water metabolism working group.[5] Some of its participants attended a workshop where the Organic Data Science framework was being used. They found the framework useful and are just starting to train new users. They decided to start a new top-level task within the age of water site, since they intend to share some datasets and software with the age of water research group. This group poses new challenges in terms of maintaining their identity while being part of a larger site with many other activities that are irrelevant to them.
- The ENIGMA consortium for neuroimaging genetics.[6] This consortium includes more than 70 institutions that collaborate to do joint neuroscience studies. The institutions keep their data locally, but they all agree to the method and software to

[5] http://www.gleon.org/research/working-groups/lake-metabolism.

[6] http://enigma.ini.usc.edu/publications/the-enigma-consortium-in-review/.

be used to analyze their data. They organize themselves into working groups, each group studies a particular disease (e.g., autism) and cohort (e.g., children). A major driver for them is to use the Organic Data Science wiki to track what institutions participate in what study, the characteristics of their datasets, and the point person in that institution for each particular study. The Task Cloning capability is particularly useful here, as is the description of data and workflows. A requirement of this group is that some information needs to remain private.

- The GPF group publishing a special issue of a journal. This is a group of geoscientists preparing articles that follow a similar format in that they publish explicitly all datasets, software, and workflows used to generate the results in the paper. The site is being used to coordinate the activities involved in tracking the status of each paper, and to compare the approaches in different papers.
- The iSamples collaboration for cataloging field science samples in geosciences. There are many catalogs of geosciences samples, and this group wishes to create a meta-catalog that will enable scientists to find an appropriate catalog to deposit their samples. An interesting challenge in this collaboration is creating structured descriptions of the curation procedures for each catalog.

All of the above collaborations are in their initial stages. Each collaboration has chosen a few selected people who have started to populate their site. Each also has specific extensions or customizations that they would like to see in the wiki. Many of the extensions described in Sect. 3 are useful for several of these wikis.

Table 1 shows some data for scientific collaboration groups. The vast majority of defined RDF-Triples are pre-defined metadata properties, as we have not yet emphasized the creation of dynamically defined properties when we train new users. As the site grows in content, we expect that these properties will be most useful in organizing information and exposing other thematic dimensions for tasks, people, and other resources in the site. At the moment, we have only trained a few selected users to create new semantic properties.

**Table 1.** Highlights of communities using the Organic Data Science framework.

| Community | # Pages | # Tasks | # Tasks with completed metadata | Avg. of task completion rates | # Registered users | # RDF triples |
|---|---|---|---|---|---|---|
| Age of water | 759 | 380 | 350 | 43.95 % | 53 | 2475 |
| ENIGMA | 204 | 80 | 2 | 2.50 % | 6 | 299 |
| GPF | 239 | 168 | 168 | 26.19 % | 32 | 1536 |
| ODS Framework | 417 | 77 | 61 | 77.92 % | 19 | 681 |
| ODS training | 1,235 | 1115 | 1112 | 99.64 % | 36 | 9219 |

We hope to create an ecosystem of developers of the Organic Data Science framework that will contribute to the existing extensions, create new ones, and share their codes so further collaborations can customize the design of their sites.

## 5 Related Work

Bry et al. [6] give a detailed overview of semantic wikis and a thorough comparison of semantic wiki frameworks. Semantic wikis have been used for scientific collaborations. Most of them are used to track particular entities, such as genes[7] or mutations.[8] The CSDMS wiki[9] is used to describe science software, and could be integrated as an external software repository.

Shared workflow repositories (e.g., [7]) allow scientists to collaborate by reusing computational methods, but do not include semantic properties for workflow tasks.

Some Web collaboration tools are also centered on coordinating tasks. For example, software development tools support issue tracking and task formulation. A major difference is that our framework is driven by science goals from the start, where each task addresses some aspect of a science goal and can result in scientific objects (software, datasets, etc.) that can be described as semantic objects in their own right.

## 6 Conclusions

This paper has presented the Organic Data Science framework, a new approach for scientific collaboration that opens the science process and exposes information about shared tasks, participants, and other relevant entities. The framework enables scientists to formulate new tasks and contribute to tasks posed by others. The framework is currently in use by a community, and is beginning to be used by others.

There are many areas of future work. Setting up the framework for new communities is a non-trivial process. The software installation is easy, but a site has to be carefully managed to jumpstart the contributions. We are investigating mechanisms to document this process and facilitate the initial stages. We continue to explore different requirements for supporting scientific collaborations in a variety of contexts.

**Acknowledgments.** We gratefully acknowledge funding from the US National Science Foundation under grant IIS-1344272.

## References

1. Krötzsch, M., Vrandecic, D., Völkel, M., Haller, H., Studer, R.: Semantic Wikipedia. J. Web Semant. **5**(4), 251–261 (2007)
2. Introne, J., Laubacher, R., Olson, G.M., Malone, T.W.: Solving wicked social problems with socio-computational systems. KI-Künstliche Intell. **27**(1), 45–52 (2013)
3. Kraut, R.E., Resnick, P.: Building Successful Online Communities: Evidence-Based Social Design. MIT Press, Cambridge (2011)

[7] http://en.wikipedia.org/wiki/Portal:Gene_Wiki.

[8] http://www.snpedia.com/.

[9] http://csdms.colorado.edu/.

4. Michel, F., Gil, Y., Hauder, M.: A virtual crowdsourcing community for open collaboration in science processes. In: Americas Conference on Information Systems (AMCIS) (2015)
5. Garijo, D., Gil, Y., Corcho, O.: Towards workflow ecosystems through semantic and standard representations. In: Proceedings of the Ninth Workshop on Workflows in Support of Large-Scale Science (WORKS), held in conjunction with the IEEE ACM International Conference on High-Performance Computing (SC) (2014)
6. Bry, F., Schaffert, S., Vrandečić, D., Weiand, K.: Semantic Wikis: approaches, applications, and perspectives. In: Eiter, T., Krennwallner, T. (eds.) Reasoning Web 2012. LNCS, vol. 7487, pp. 329–369. Springer, Heidelberg (2012)
7. De Roure, D., Goble, C.A., Stevens, R.: The design and realisation of the myExperiment virtual research environment for social sharing of workflows. Future Gener. Comput. Syst. **25** (5), 561–567 (2009)