A Virtual Crowdsourcing Community for Open Collaboration in Science Processes

Full Paper

Felix Michel
Technical University Munich
Boltzmannstr. 3
Munich, BY 85748, DE
felix.michel@tum.de

Yolanda Gil
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292, US
gil@isi.edu

Varun Ratnakar
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292, US
varunr@isi.edu

Matheus Hauder
Technical University Munich
Boltzmannstr. 3
Munich, BY 85748, DE
matheus.hauder@tum.de

Abstract

Although science has become an increasingly collaborative endeavor over the last hundred years, only little attention has been devoted to supporting scientific communities. Our work focuses on scientific collaborations that revolve around complex science questions that require significant coordination to synthesize multi-disciplinary findings, enticing contributors to remain engaged for extended periods of time, and continuous growth to accommodate new contributors as needed as the work evolves over time. This paper presents a virtual crowdsourcing community for open collaboration in science processes to address these challenges. Our solution is based on the Semantic MediaWiki and extends it with new features for scientific collaboration. We present preliminary results from the usage of the interface in a pilot research project.

Keywords

Organic Data Science, Semantic MediaWiki, scientific collaboration, virtual crowdsourcing community.

Motivation

Over the last hundred years, science has become an increasingly collaborative endeavor. Scientific collaborations, sometimes referred to as “collaboratories” and “virtual organizations”, range from those that work closely together and others that are more loosely coordinated (Bos et al. 2007; Ribes and Finholt 2009). 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), others form around a shared software base (e.g., SciPy), and others around a shared scientific question (e.g., the Human Genome Project). The application of crowdsourcing approaches in these collaboratories through virtual communities provides manifold new opportunities to realize the potential of collective intelligence in science. Scientists with diverse knowledge and skills around the globe could be accessed by opening scientific processes that expose all tasks and activities publicly to achieve a shared scientific question.

Our work focuses on scientific collaborations that are driven by a shared scientific question and require the successful integration of ideas, models, software, data, other resources as well as scientists from different disciplines. For all these reasons, even though such scientific collaborations do occur the potential of crowdsourcing through virtual communities in science has not been fully uncovered. Yet, virtual crowdsourcing communities are needed to address major engineering and science challenges in our future (e.g., (NAE 2014)). The approach in this paper integrates findings on successful communities from social sciences and crowdsourcing processes to facilitate open collaboration in science.

Crowdsourcing can be defined as a transformation of tasks that are traditionally performed by employees to the crowd through an open call. Crowdsourcing models consist of an initiator who crowdsources a task, a mediating platform, and contributors from the crowd that perform these tasks. In science incorporating contributors from the crowd could provide valuable knowledge and resources to elaborate complex scientific questions. Initiators who crowdsource tasks might be employees within a research project or organization that have only limited resources or require specific skills from external experts. In general crowdsourcing processes can be described along four dimensions (Geiger et al. 2011). In the first dimension, preselection of contributors is concerned with restrictions on the group of potential contributors which might be qualification-based, context-specific or a combination of both values. In the second dimension the accessibility of peer contributions describes to what extent contributors are able to access the contributions of others. Possible values for this dimension range from modify, assess, view or no access rights allowed. In the third dimension, two options for the aggregation of contributions are possible, i.e. contributions can be integrated or the best solutions are chosen selectively. In the fourth dimension, the remuneration for contributions can be distinguished with a fixed amount, an amount based on success or no remuneration.

This paper presents the Organic Data Science framework as a virtual crowdsourcing community to support scientific collaborations and processes that revolve around complex science questions that require significant coordination to synthesize multi-disciplinary findings, enticing contributors to remain engaged for extended periods of time, and continuous growth to accommodate new contributors as needed as the work evolves over time. Regarding the dimensions of crowdsourcing processes the Organic Data Science framework can be considered as an integrative sourcing without remuneration. In our approach contributors need to have a scientific background that matches with the requirements of the research project. Members in the scientific community have access to contributions of others to make changes similar to e.g. Wikipedia or OpenStreetMap. Results of scientific communities in the Organic Data Science framework are integrated and contribute to a shared scientific question. Although the framework provides no remuneration for contributing scientists, our framework incorporates social design principles from successful online communities to leverage motivation and commitment of contributors. To the best of our knowledge the Organic Data Science framework is the first approach that aims to develop an integrative sourcing without remuneration for scientific crowdsourcing.

For this purpose the framework identifies a set of features that help to retain and engage scientists in their communities. This innovative set of features is based upon social design principles from successful online communities. Our organic data science framework incorporates these features in the design of a collaborative user interface that supports: 1) *self-organization of the community* through user-driven dynamic task decomposition, 2) *on-line community support* by incorporating social design principles and best practices, 3) *an open science process* by capturing new kinds of metadata about the collaboration that provide invaluable context to newcomers. Users formulate science tasks to describe the what, who, when, and how of the smaller activities pursued within the collaboration on an overarching shared scientific question. The interface is designed to entice contributors to participate and continue involved in the specific tasks they are interested in. The framework is in its early stages of development, and it evolves to accommodate user feedback and to incorporate new collaboration features.

The paper continues with an overview of the state of the art in collaboration in scientific processes. We then introduce our approach to support task-oriented self-organizing communities for open scientific collaborations. Based on this approach we present our implementation that is based on the Semantic Media Wiki platform which we extended with new features to allow open scientific processes in virtual communities. The new features are mapped against social design principles that are retrieved from current literature and patterns that we collected from existing successful communities. Preliminary evaluation data are presented from the usage of the prototype with scientists working on a pilot research project that focuses on theoretical and experimental aspects of the isotopic “age” of water in watershed-lake systems which requires a community that incrementally grows with unanticipated scientists that bring in the required skills and resources. Regarding our research hypotheses of retaining scientists in their communities through novel features that are based on social design principles the evaluation investigates the following three research questions: Is the framework helping users to organize their work? Is the framework helping to create communities? Is the framework helping to open science processes?

Related Work

We find inspiration in the Polymath project, set up to collaboratively develop proofs for mathematical theorems (Gowers and Nielsen 2009; Nielsen 2012), where professional mathematicians collaborate with volunteers that range from high-school teachers to engineers to solve mathematics conjectures. It uses common Web infrastructure for collaboration, interlinking public blogs for publishing problems and associated discussion threads (Nielsen 2012) with wiki pages that are used for write-ups of basic definitions, proof steps, and overall final publication (Gowers and Nielsen 2009). Interactions among contributors to share tasks and discuss ideas are regulated by a simple set of social norms for the collaboration (Gowers and Nielsen 2009). The growth of the community is driven by the tasks that are posted, as tasks are decomposed into small enough chunks that contributors can take on.

Another project that has exposed best practices of a large collaboration is ENCODE (Birney 2012; Encode 2004). In ENCODE, the tasks that are carved out for each group in the collaboration are formally assigned since there is funding allocated to the tasks. In addition the collaboration members are selected beforehand. Despite these differences with our project, we share the explicit assignment of tasks in service of science goals. Figure 2 outlines the best practices and lessons learned from these two projects that are applicable to our work. There have been many studies of on-line communities (Kraut and Resnick 2011), notably on Wikipedia. Our work builds on the social design principles uncovered by this research. However, our belief is that scientific work is best organized around tasks, not topic pages. An analysis of Wikipedia shows a continuously increasing readership and a decreasing contribution since 2007, pointing to the need to better coordinate work (Morgan et al. 2014).

A study on Electronic Lab Notebooks shows the benefits of structuring knowledge in an ad-hoc and simple manner (Oleksik et al. 2014). Other studies have demonstrated the benefits of using a shared communication board to facilitate collaborative decision making for patient care (Kane et al. 2013). A study of MathOverflow shows how the quality of answers can be improved collaboratively (Tausczik et al. 2014). Collaborative user interfaces that have been used in science include semantic wikis (e.g., (Huss et al. 2010)), workflow repositories (De Roure et al. 2009), and argumentation systems (e.g., (Introne et al. 2013)). However, their adoption remains limited. In contrast, popular collaborative Web frameworks are widely used in science, including code repositories, blogs, and wikis. For example, issue tracking tools are popular to coordinate programmer teams, and can be used for managing other kinds of tasks. Our approach shares some important features with these tools in tracking tasks. However, our approach is better positioned to address social issues such as incentives, motivation, and enticing newcomers.

(Ribes and Finholt 2009) analyze the challenges of organizing work in four scientific collaborations: GEON (Geosciences Network), LEAD (Linked Environments for Atmospheric Discovery), WATERS (Water and Environmental Research Systems), and LTER (Long-Term Ecological Research). They found that major challenges for organizing work were: 1) the tension between planned work, with its work breakdown structures with deadlines, versus emergent organization as new requirements and unknowns are uncovered, 2) the tradeoff that participants face between doing basic research and contributing to the technical development in support of the research, and 3) the desire to incorporate innovations while needing a stable framework to do research. Other studies have uncovered similar needs (Steinhardt and Jackson 2014). Organic Data Science is poised to offer the flexibility of easily incorporating emergent tasks and people, and the enticement to participants through acknowledgement of contributions so that uneven support from particular contributors is properly exposed.

The coordination of work has been a focus of formal theories (e.g., SharedPlans (Lochbaum et al. 1990)) and practical implementations of those theories (Rich et al. 2005; Rich et al. 2001). The work has focused either on human-computer dialogue or multi-agent coordination. In our case, the coordination is among humans. A promising area of future work is to investigate if these collaboration theories and frameworks could be incorporated into the design of our multi-human collaboration interface. Task-oriented interfaces have been developed for scientific computing, where data analysis tasks are cast as workflows whose validation and execution are managed by the system (Chin et al. 2002). In our framework, tasks can be decomposed into more and more specific and well-defined tasks that can be turned into workflows that can be executed for data analysis. The interface between our framework and workflows is an area of planned work.

Approach

We are developing an Organic Data Science framework to support task-oriented self-organizing on-line communities for open scientific collaboration. Its key features are:

1. **Self-organization of the work**, through an interface that supports scientists to organize joint tasks and to easily track where they can contribute and when
2. **Sustainable on-line communities**, through an interface that incorporates principles from social sciences research on successful on-line collaborations, including best practices for retention and growth of the community
3. **Open science processes that expose all tasks and activities publicly**, through an interface that captures new kinds of metadata about the collaboration so all participants (especially newcomers) can immediately catch up with the work being done

Our goal is to reduce the coordination effort required and to lower the barriers to growing the community.

Self-Organization

Our approach is to use tasks as an organizational mechanism for coordination, and to allow users to create joint tasks, decompose them into smaller subtasks, and easily track their status. Tasks can be seen as a shared tool for social cognition (Hutchins 1995), which considers that in collaborative settings the expertise is not only in the minds of individuals but in the organization of the tools and objects that they share. Processes and tasks have been shown to be a key to collaboration in science laboratories (Chandrasekaran and Nernessian 2014), to coordinate work in multi-agent systems (Grosz and Sidner 1988), and to the productivity of knowledge workers in an organization (Davenport 2013).

Decomposition of subtasks is an important aspect of describing tasks. Many explanations of procedures, including scientific and technical expositions, exhibit goal-oriented hierarchical structure (Britt and Larson 03). The temporal relations among subtasks are also important (Pietras and Coury 1994). The user interface should be designed so users have some initial structure to express tasks. (Van Merrienboer 2003) proposes the use of process worksheets to guide students through complex tasks. (Mahling and Croft 1993) also found that the formulation of tasks is greatly improved through form-based interfaces.

Sustainable On-Line Communities

Our approach is to form and sustain communities around science goals, not simple collaborations. Numerous studies about successful on-line communities provide useful design principles for our framework (Kraut and Resnick 2011), with topics as varied as the design of the editorial process (Spinellis and Louridas 2008), community composition and activities (Gil and Ratnakar 2013), incentives to contributors (Mao et al. 2013), critical mass of contributors (Raban et al. 2010), coordination (Kittur et al. 2009), group composition (Lam et al. 2010), conflict (Kittur et al. 2010), trust (McGuinness et al. 2006), and user interaction design (Hoffman et al. 2009). Figure 1 summarizes the social principles that we are using in our approach. We follow the organization used in (Kraut and Resnick 2011), focusing in this paper on social principles that are relevant to early stages of the community. In the next section we explain how they map to features in our user interface.

Opening Science Process

Our approach is to make the collaborative science processes explicit, so that everyone can examine the status of the collaboration and access the rationale of the current activities being pursued. These collaborative processes may be explicitly articulated but are never captured. (Polanyi and Sen 1967) coined the terms and discussed differences between tacit and explicit knowledge of individuals in organizations. While explicit knowledge can be communicated in formal languages that can be processed by other individuals, people have tacit knowledge that they cannot explicitly express. In their theory on organizational knowledge creation, Nonaka and Takeuchi described the transformation modes between tacit and explicit knowledge with socialization, externalization, internalization, and combination (Takeuchi and Nonaka 2004; Nonaka and Takeuchi 1995). We aim at externalizing the tacit knowledge of scientists about the collaboration itself.

A. Starting communities
- A1. Carve a niche of interest, scoped in terms of topics, members, activities, and purpose
- A2. Relate to competing sites, integrate content
- A3. Organize content, people, and activities into subspaces once there is enough activity
- A4. Highlight more active tasks
- A5. Inactive tasks should have "expected active times"
- A6. Create mechanisms to match people to activities

B. Encouraging contributions through motivation
- B1. Make it easy to see and track needed contributions
- B2. Ask specific people on tasks of interest to them
- B3. Simple tasks with challenging goals are easier to comply with
- B4. Specify deadlines for tasks, while leaving people in control
- B5. Give frequent feedback specific to the goals
- B6. Requests coming from leaders lead to more contributions
- B7. Stress benefits of contribution
- B8. Give (small, intangible) rewards tied to performance (not just for signing up)
- B9. Publicize that others have complied with requests
- B10. People are more willing to contribute: 1) when group is small, 2) when committed to the group, 3) when their contributions are unique

C. Encouraging commitment
- C1. Cluster members to help them identify with the community
- C2. Give subgroups a name and a tagline
- C3. Put subgroups in the context of a larger group
- C4. Make community goals and purpose explicit
- C5. Interdependent tasks increase commitment and reduce conflict

D. Dealing with newcomers
- D1. Members recruiting colleagues is most effective
- D2. Appoint people responsible for immediate friendly interactions
- D3. Introducing newcomers to members increases interactions
- D4. Entry barriers for newcomers help screen for commitment
- D5. When small, acknowledge each new member
- D6. Advertise members particularly community leaders, include pictures
- D7. Provide concrete incentives to early members
- D8. Design common learning experiences for newcomers
- D9. Design clear sequence of stages to newcomers
- D10. Newcomers go through experiences to learn community rules
- D11. Provide sandboxes for newcomers while they are learning
- D12. Progressive access controls reduce harm while learning

Figure 1. Selected social principles from (Kraut and Resnick 2012) for building successful online communities that can be applied to Organic Data Science. We focus on social principles that are relevant to early stages of the community, and leave out more advanced principles (e.g., for retention of members and for regulating behavior).

E. Best practices from Polymath
- E1. Permanent URLs for posts and comments, so others can refer to them
- E2. Appoint a volunteer to summarize periodically
- E3. Appoint a volunteer to answer questions from newcomers
- E4. Low barrier of entry: make it VERY easy to comment
- E5. Advance notice of tasks that are anticipated
- E6. Keep few tasks active at any given time, helps focus

F. Lessons learned from ENCODE
- F1. Spine of leadership, including a few leading scientists and 1-2 operational project managers, that resolves complex scientific and social problems and has transparent decision making
- F2. Written and publicly accessible rules to transfer work between groups, to assign credit when papers are published, to present the work
- F3. Quality inspection with visibility into intermediate steps
- F4. Export of data and results, integration with existing standards

Figure 2. Selected best practices from the Polymath (Nielsen 2012) project and lessons learned from ENCODE (Encode 2004) that can be applied to the initial design of our Organic Data Science framework.

The Organic Data Science Wiki[1]

In this section we describe the Organic Data Science Wiki (ODSW), our current implementation of the Organic Data Science framework. It is built as an extension of the Semantic Media Wiki platform (Bry et al. 2012; Krötzsch et al. 2011), and uses its semantic capabilities to structure the content of the site, including task properties and user properties. The semantic wiki provides an intuitive user interface that hides from users any formal semantic notation (Bry et al. 2012; Gil 2013). We highlight here major features of the interface that implement our approach. These features are summarized in Table 1.

Self-Organization through User-Driven Dynamic Task Decomposition

ODSW allows users to create tasks, describe them, and decompose them into smaller subtasks. 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. Figure 3 shows a screenshot of a task page, with the features highlighted in blue circles. Task pages follow a pre-defined structure that is automatically presented to the user when a new task is created (❶). Subtasks can be added to form a hierarchical task structure (❸,❺). Users are asked to specify metadata (❷a), which are major properties of the task such as start and target dates. These metadata properties enable the ODSW to assist users to manage tasks by generating timelines (❻), a task state summarizing subtask progress (❿), and alerts when they are late (❼). Users can sign up for a task either as "owners", which makes them responsible for the task getting done, or "participants", which means they will contribute to the task.

We distinguish between several categories of metadata. *Pre-defined metadata* are properties of tasks that ODSW will use to assist users to manage tasks (❹,❼,❾,❿). Pre-defined metadata can be *required* or *optional*. Required metadata includes the start date, target date, task owner, and task type. Tasks whose required metadata is incomplete have special status in ODSW 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. *Dynamically-defined metadata* (❷b) 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.

An important required metadata property is the task type (high, medium, and low level). The progress to date for low-level tasks is provided manually by their owners or participants, since the tasks have small duration. ODSW calculates the progress of higher-level tasks based on their subtasks and their start as well as target dates. 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. High-level task are colored in lighter green and lower-level tasks in darker green.

Sustainable On-Line Communities through Best Practices

The user interface of ODSW is designed to support the formation of an on-line community and its growth. We follow the successful social design principles (see Figure 1) and combine it with best practices from projects such as Polymath and ENCODE (see Figure 2). We highlight here how the interface is designed to address some of these social principles.

Several social principles (A1-A6) address the formation of the community. They are most noticeable in the main page of the site (see Appendix). It describes clearly the science and technical objectives of the project, displays a summary of currently active tasks, and shows the leadership and major contributors (❿). In geosciences, the models used in the project are important to anchor the work for newcomers, so they are also shown in the main page. ODSW automatically generates the model and contributor tables from the current contents with a semantic wiki query. Those tables highlight properties of note, which allow newcomers to match ongoing work to their personal interests. Dealing with newcomers is another important aspect of creating an on-line community. Social principles D1-D12 address this and we set up a separate site to train new users (see Appendix). This training site also uses ODSW.

[1] http://www.organicdatascience.org/ageofwater/index.php/Main_Page, last accessed on: 2015-02-17

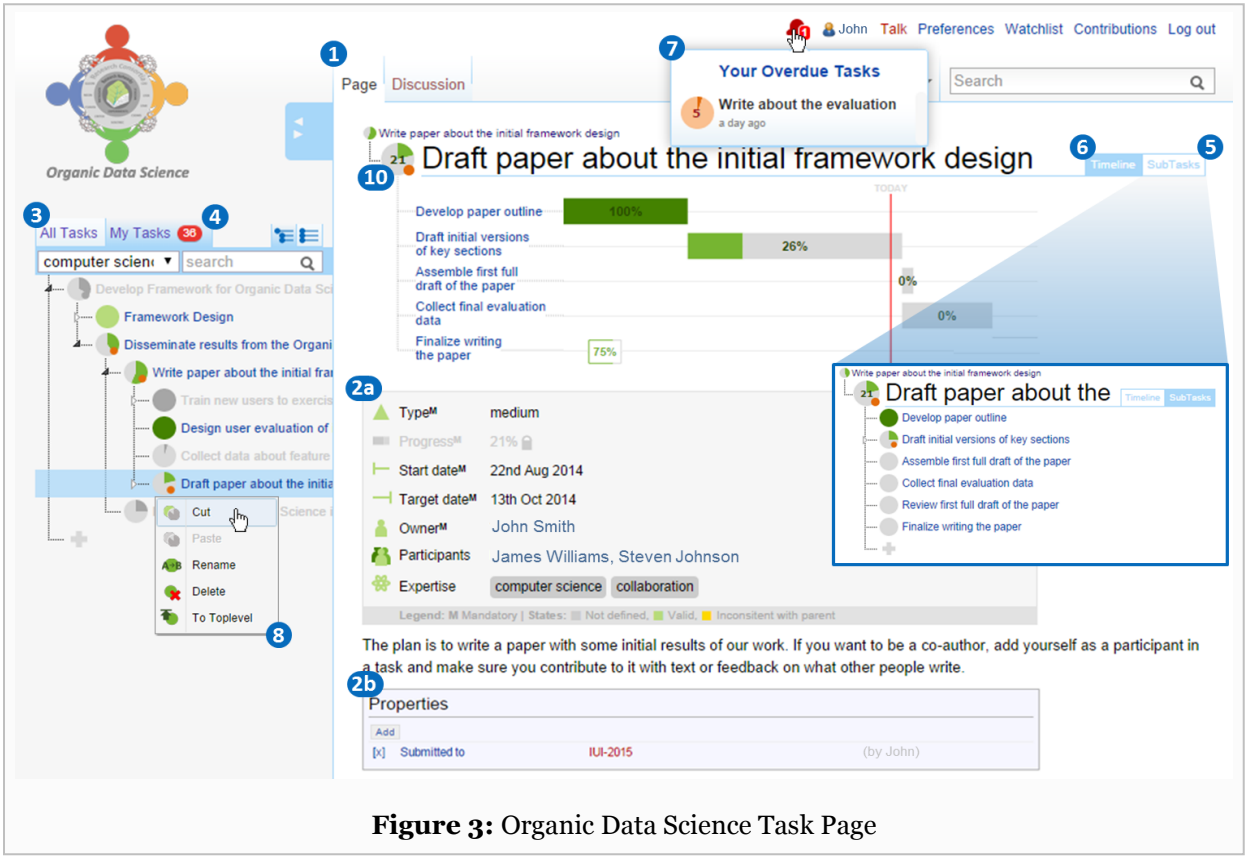


Figure 3: Organic Data Science Task Page

Feature	*Feature Description*	*Objectives* I	II	III	*Social Principles*
⓪ **Welcome Page**	Describes clearly the science and technical project objectives summarizes currently active tasks, and shows lead contributions.		✓	✓	A1, A2, A3, B7, D1, D5, D6, D7, E2, E6, F1, F2, E4
❶ **Task Representation**	Tasks have a unique identifier (URL), and are organized in a hierarchical subtask decomposition structure.	✓		✓	A3, A4, A6, B1, B3, B10, C2, C3, C4, C5, E1, E5, F3
❷ **Task Metadata**	a) Task metadata are properties, such as start date and target date. b) User structured properties. All metadata is stored as semantic properties.	✓	✓	✓	A4, A5, A6, B1, B2, B4, B5, B6, C1, C2, C5, F3
❸ **Task Navigation**	Tasks can expand until a leaf task is reached. Additionally users can search for task titles and apply an expertise filter.	✓		✓	B1, B4, B10, C1, C2, C3, C4, C5, F3
❹ **Personal Worklist**	The worklist contains the subset of tasks from the task navigation for which the user is owner or a participant.	✓	✓	✓	A4, B1, B4, C3
❺ **Subtask Navigation**	Subtasks of the currently opened task are presented.	✓		✓	B1, B5, B9, B10 C5, F3
❻ **Timeline Navigation**	All subtasks are represented based on their start, target times, and completion status in a visualization based on a Gantt chart.	✓		✓	A4, A5, B1, B5, E5, F3
❼ **Task Alert**	Signals when a task is not completed and the target date passed	✓	✓		B1, B4
❽ **Task Management**	The interface supports creating, renaming, moving and deleting tasks. For usability reasons, all these actions can be reversed.	✓			A3, B3, B10, F3
❾ **User Tasks and Expertise**	The interface allows users to easily see what others are working on or have done in the past. This creates a transparent process. (see Appendix)	✓	✓	✓	B1, B2, B5, B8, B10, C1, C5
❿ **Task State**	Small icons visualize the state of each task intuitively.	✓	✓	✓	B1, B5, E5
⓫ **Training New Members**	A separate site is used 1to train new users in a sandbox environment, where training tasks are explicit. (see Appendix)		✓	✓	D2, D3, D4, D8, D9, D10, D11, D12, E3, E4

Table 1. Mapping between features in the Organic Data Science Wiki and social principles

Opening Science Processes through Explicit Metadata Capture

ODSW creates a transparent work process. Anyone can see the contents of the site, the process being followed by the whole community, and the tasks being undertaken by different subgroups are open and accessible. In order to edit the contents, users have to become contributors by getting a login and undergoing training (⓫). Decomposing complex tasks into smaller manageable tasks also makes the science process more transparent. The ODSW interface allows to drill down into subtasks or drill up to the more general parent task (❸,❹,❺,❻,❽). Users who own small tasks can see the context and importance of their tasks.

Defining explicit task metadata (❷) such as a task type, progress, owner, participants or start and target date helps to make the process more transparent for other users. The interface exploits this metadata to help users find what is relevant to them. For example, ODSW groups the tasks for which the user is owner or a participant and forms personal task lists (❹,❾). This allows users to easily see what others are planning to work on or have worked on in the past. Another example is that hovering over a certain expertise value (e.g., "nutrients") fades out all tasks in the page that are not associated with that expertise. This helps new contributors find out relevant activities through the tasks that their colleagues are involved with. ODSW aggregates information from the required metadata properties and automatically generates visual task states as colored pie chart task icons (⓾).

Evaluation

We present an evaluation of our current implementation of the Organic Data Science framework. The site has been active since January 2014 and has been in use while new features were rolled out. The lead users have live discussions at biweekly telecons to discuss the design of the framework and the overall progress of the work, with all the resulting tasks captured in the wiki. We instrumented the system and started to collect data at 1st October once all the features described above were rolled out. In the first 10 weeks until 10th December we collected around 19,000 log entries, which we used for the evaluation presented here.

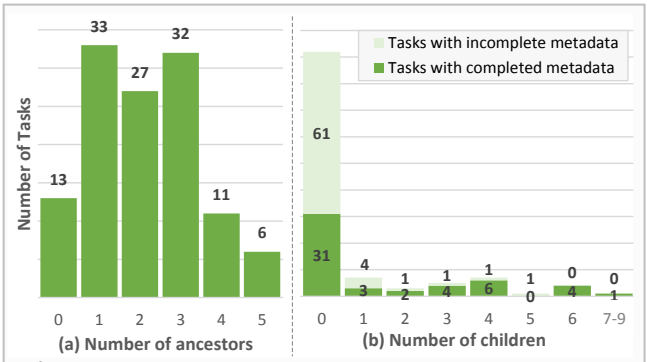


Figure 5: Subtask Hierarchies.

Is the Framework Helping Users Organize their Work?

The site contains 122 tasks. In the 10-week time period all task pages together were accessed more than 2,900 times. Person pages were accessed 328 times in total. The tasks in the current ODSW site include: 1) tasks about the science of the age of water, 2) tasks about the development of the Organic Data Science approach and its implementation, 3) tasks about outreach such as an upcoming workshop about ODSW at the annual GLEON meeting. We organized and wrote this paper collaboratively using ODSW.

Figure 5 shows data about the task hierarchies in terms of the depth (number of ancestors of tasks) and breadth (number of children). There are 13 top-level tasks, and the majority of tasks are at the next three levels of decomposition. As far as breadth, most tasks have no subtasks, and are

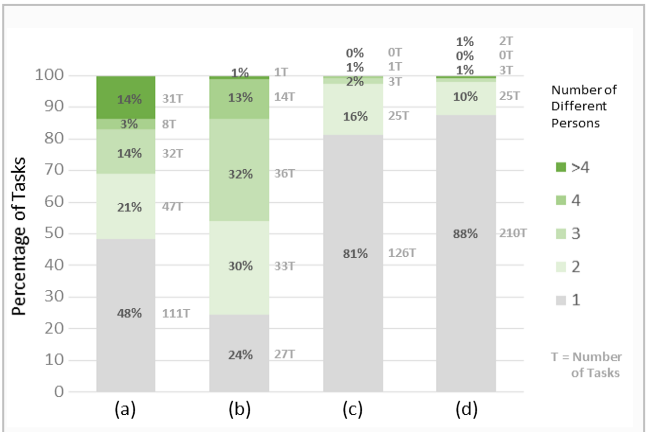


Figure 6: Task Collaboration Evaluation.

either tasks small enough that they do not require further decomposition or tasks that will take place in the future and have not yet been fleshed out. Many tasks do have several subtasks.

Is the Framework Helping to Create Communities?

We analyzed the logs to determine how many users were connecting in some way through the tasks in the site. We removed tasks with no participants, since they were created recently and did not even have an owner. We did not filter out data for tasks that were renamed or deleted.

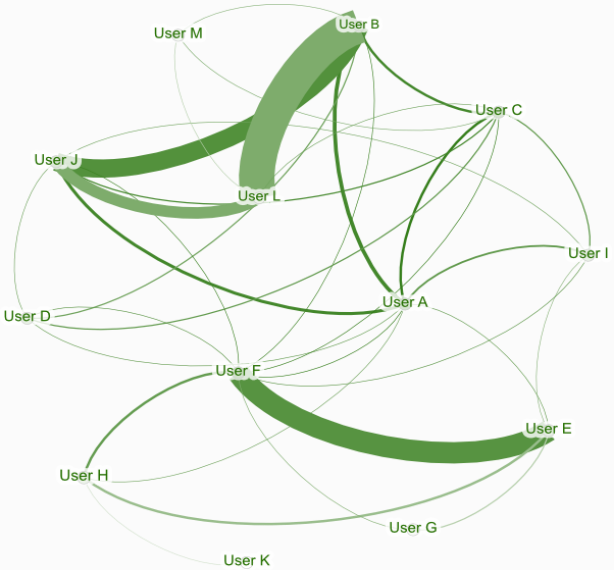


Figure 7: Organic Data Science Collaboration Graph.

Figure 6(a) shows that 52% of the tasks are visited by two or more persons. Currently 48% of all task pages are accessed by only one person. This is a high number, but we believe that this is due to the many tasks that are planned but not yet worked on since the project is still in its first year. We expect this percentage to decrease as the project progresses, particularly as it gets closer to completion. Figure 6(b) shows the total persons involved in tasks, including the participants and the owner. 72% of the tasks have two or more persons involved, and 46% have three or more. This is quite a high number of people sharing tasks. According to Figure 6(c) 81% of all tasks have their metadata edited by only one person. This is expected, since typically the task owner adds the initial metadata. But 19% of the tasks have their metadata edited by two or more persons. This indicates that non-owners have an interest in the management of the tasks. This is shown in Figure 6(d). 11% of the tasks have their content edited by two or more persons. The vast majority of the tasks have their content edited by just one person. This is a very low number, and we hope it will increase as more tasks are worked on and accomplished.

We created a network by using task metadata properties about owners and participants in tasks. Users are represented as nodes in the network, and each edge between two nodes represents that the two users are signed up for the same task one or more times. The number of tasks they have in common is expressed by the strength of edges. The result is illustrated in Figure 7. One interesting observation is that there are edges among most of the existing users, indicating collaboration activities across all participants. There are two major connected components in the graph, which are apparent at the top and the bottom of the network, indicating two strong collaboration communities. Users developing the ODSW software are at the bottom, while users working on the age of water are at the top.

Is the Framework Helping to Open the Science Processes?

This aspect of our approach is hard to evaluate, particularly since the community is still small. New users report informally that it is easy to browse the wiki and understand what tasks are currently active, why are they being pursued, who is involved, and what their scope and goals are. In the future, we plan to conduct surveys with users about the utility of the framework to help them understand the status of the collaboration.

The site initially had four users, who started to create content and tasks. So far the community has grown by direct referral (per principle D1 of Figure 1). Within 3 months a handful of additional users were brought in to help with specific tasks. In the last few weeks, a few more have been added. The site currently has 18 registered users, which include computer scientists, hydrologists, ecologists, limnologists, and geoinformaticians. At the end of October, a first outreach invitation-only workshop was held at the GLEON annual meeting which had about 40 registered participants interested in the age of water. All workshop activities are being managed using ODSW by the organizers and by the participants themselves, and we expect some fraction of them to remain involved. So far new users have been added painlessly.

New users report that the training tasks take around one hour. Our logs show that they did not access the documentation once training was completed. New users are creating tasks and participating in them, and the logs also show they are not undoing any of these actions.

Conclusion

We have presented the Organic Data Science Framework that provides a novel set of features in the Semantic MediaWiki that help to retain scientists in their communities. The main features of this framework are a task-centered organization, the incorporation of social design principles, and the open exposure of scientific processes.

We continue to collect data about the on-line activities in several scientific communities that use the Organic Data Science Wiki. We have specific hypotheses about how the maturity of the project will affect the management of tasks, about how the growth of the communities will affect the amount of on-line coordination that occurs, and about the task structure as the scope of the work increases. Future work includes analyzing the evolution of the communities in quantitative terms.

Acknowledgements

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

REFERENCES

Begley, C. Glenn, and Lee M. Ellis 2012. "Drug development: Raise standards for preclinical cancer research," *Nature* 483.7391: 531-533.

Birney, Ewan 2012. "The making of ENCODE: lessons for big-data projects," Nature489.7414: 49-51.

Board, M. E. A. (2005). "Ecosystems and human well-being: synthesis," Washington, DC: World Resources Institute.

Bos, N., Zimmerman, A., Olson, J., Yew, J., Yerkie, J., Dahl, E., & Olson, G. 2007. "From shared databases to communities of practice: A taxonomy of collaboratories," *Journal of Computer-Mediated Communication*, 12(2), 652-672.

Brickley, D., and Guha, R. V. 2014. "RDF Vocabulary Description Language 1.0: RDF Schema," Available from http://www.w3.org/TR/2014/REC-rdf-schema-20140225/

Britt, M. A., and Larson, A. A., 2003. "Constructing Representations of Arguments," *Journal of Memory and Language*, 48.

Bry, F., Schaffert, S., Vrandečić, D., and Weiand, K. 2012. „Semantic wikis: Approaches, applications, and perspectives," (pp. 329-369). Springer Berlin Heidelberg.

Chandrasekharan, S., and Nersessian, N. J. 2014. "Building Cognition: The Construction of Computational Representations for Scientific Discovery," *Cognitive science*.

Chin Jr, G., Leung, L. R., Schuchardt, K. and Gracio, D. 2002. "New paradigms in problem solving environments for scientific computing." *In Proceedings of the 7th international conference on Intelligent user interfaces* (pp. 39-46). ACM.

Davenport, T. H. 2013. "*Thinking for a living: how to get better performances and results from knowledge workers.*" Harvard Business Press.

De Roure, D., Goble, C., and Stevens, R. 2009. "The design and realization of the Virtual Research Environment for social sharing of workflows." *Future Generation Computer Systems*, 25(5), 561-567.

ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696), 636-640.

Garijo, D., Kinnings, S., Xie, L., Xie, L., Zhang, Y., Bourne, P. E., and Gil, Y. 2013. "Quantifying reproducibility in computational biology: the case of the tuberculosis drugome." *PloS one*, 8(11), e80278.

Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., and Schader, M. (2011). „Managing the Crowd: Towards a Taxonomy of Crowdsourcing Processes." In Seventeenth Americas Conference on Information Systems, Detroit, Michigan.

Gil, Y. 2013. "Social Knowledge Collection." *In Handbook of Human Computation* (pp. 285-296). Springer New York.

Gil, Y., and Ratnakar, V. 2013. "Knowledge capture in the wild: a perspective from semantic wiki communities." *In Proceedings of the seventh international conference on Knowledge capture* (pp. 49-56). ACM.
Gowers, T., and Nielsen, M. 2009. "Massively collaborative mathematics." *Nature*, 461(7266), 879-881.
Grosz, B. J., and Sidner, C. L. 1988. "Plans for discourse" (No. BBN-6728). BBN LABS INC CAMBRIDGE MA.
Hoffmann, R., Amershi, S., Patel, K., Wu, F., Fogarty, J., and Weld, D. S. 2009. "Amplifying community content creation with mixed initiative information extraction." *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1849-1858). ACM.
Huss, J. W., Lindenbaum, P., Martone, M., Roberts, D., Pizarro, A., Valafar, F., and Su, A. I. 2009. "The Gene Wiki: community intelligence applied to human gene annotation." *Nucleic acids research*, gkp760.
Hutchins, E. 1995. "How a cockpit remembers its speeds." *Cognitive science*, 19(3), 265-288.
Introne, J., Laubacher, R., Olson, G., & Malone, T. (2013). "Solving wicked social problems with socio-computational systems." *KI-Künstliche Intelligenz*, 27 (1), 45-52.
Kane, B. T., Toussaint, P. J., & Luz, S. 2013. „Shared decision making needs a communication record." *In Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 79-90). ACM.
Kittur, A., Lee, B., & Kraut, R. E. 2009. "Coordination in collective intelligence: the role of team structure and task interdependence." *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1495-1504). ACM.
Kittur, A., and Kraut, R. E. 2010. "Beyond Wikipedia: coordination and conflict in online production groups." *In Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 215-224). ACM.
Kraut, R. E., Resnick, P., Kiesler, S., Burke, M., Chen, Y., Kittur, N., and Riedl, J. 2012. *„Building successful online communities: Evidence-based social design*." MIT Press.
Krötzsch, M., Vrandečić, D., and Völkel, M. 2006. "Semantic mediawiki." In The Semantic Web-ISWC 2006 (pp. 935-942). Springer Berlin Heidelberg.
Lam, S. K., Karim, J., & Riedl, J. 2010. "The effects of group composition on decision quality in a social production community." *In Proceedings of the 16th ACM international conference on supporting group work* (pp. 55-64). ACM.
Levin, S. A., and Clark, W. C. 2010. "Toward a science of sustainability."
Lochbaum, K. E., Grosz, B. J., and Sidner, C. L. 1990. „Models of plans to support communication: An initial report." *In AAAI* (pp. 485-490).
Mahling, D. E., and Croft, W. B. 1993. "Acquisition and support of goal-based tasks." *Knowledge acquisition*, 5 (1), 37-77.
Mao, A., Kamar, E., Chen, Y., Horvitz, E., Schwamb, M. E., Lintott, C. J., and Smith, A. M. 2013. "Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing." *In First AAAI Conference on Human Computation and Crowdsourcing*.
McGuinness, D. L., Zeng, H., Da Silva, P. P., Ding, L., Narayanan, D., and Bhaowal, M. 2006. "Investigations into Trust for Collaborative Information Repositories: A Wikipedia Case Study." *MTW*, 190.
Morgan, J. T., Gilbert, M., McDonald, D. W., and Zachry, M. 2014. „Editing beyond articles: diversity & dynamics of teamwork in open collaborations." *In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 550-563). ACM.
NAE. 2014. "NAE Grand Challenges in Engineering". National Academy of Engineering. Available from http://www.engineeringchallenges.org.
Nielsen, M. 2012. "Reinventing discovery: the new era of networked science." Princeton University Press.
Nonaka, I., and Takeuchi, H. 1995. "The knowledge-creating company: How Japanese companies create the dynamics of innovation." Oxford university press.
Nonaka, I., and Takeuchi, H. 2004. "Hitotsubashi on knowledge management." Wiley.
Oleksik, G., Milic-Frayling, N., and Jones, R. 2014. "Study of electronic lab notebook design and practices that emerged in a collaborative scientific environment." *In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 120-133). ACM.
Pietras, C. M., & Coury, B. G. 1994. "The development of cognitive models of planning for use in the design of project management systems." *International journal of human-computer studies*, 40(1), 5-30.
Polanyi, M., and Sen, A. 1967. "The tacit dimension" (p. 108). New York: Doubleday.

Raban, D. R., Moldovan, M., and Jones, Q. 2010. "An empirical study of critical mass and online community survival." *In Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 71-80). ACM.

Ribes, D., and Finholt, T. A. 2009. "The long now of technology infrastructure: articulating tensions in development." *Journal of the Association for Information Systems*, 10(5), 5.

Rich, C., Sidner, C. L., and Lesh, N. 2001. "COLLAGEN: Applying Collaborative Discourse Theory to Human-Computer Interaction." *AI Magazine 22*.

Rich, C., Sidner, C., Lesh, N., Garland, A., Booth, S., and Chimani, M. 2005. "DiamondHelp: A collaborative task guidance framework for complex devices." *In Proceedings of the National Conference on Artificial Intelligence* (Vol. 20, No. 4, p. 1700). MIT Press.

Spinellis, D., and Louridas, P. 2008. "The collaborative organization of knowledge." *Communications of the ACM*, 51(8), 68-73.

Stephanie B. Steinhardt and Steven J. Jackson. 2014. "Reconciling Rhythms: Plans and Temporal Alignment in Collaborative Scientific Work." *Computer Supported Cooperative Work and Social Computing (CSCW)*, Baltimore, Maryland.

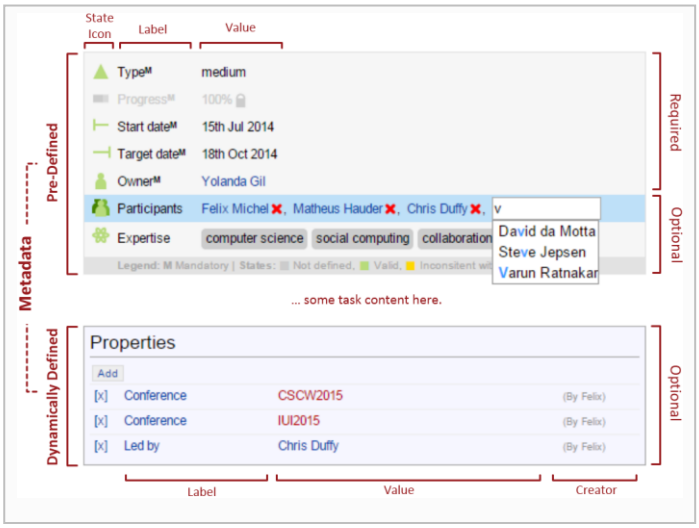
Suweis, S., Rinaldo, A., Maritan, A., and D'Odorico, P. 2013. "Water-controlled wealth of nations." *Proceedings of the National Academy of Sciences*, 110(11), 4230-4233.

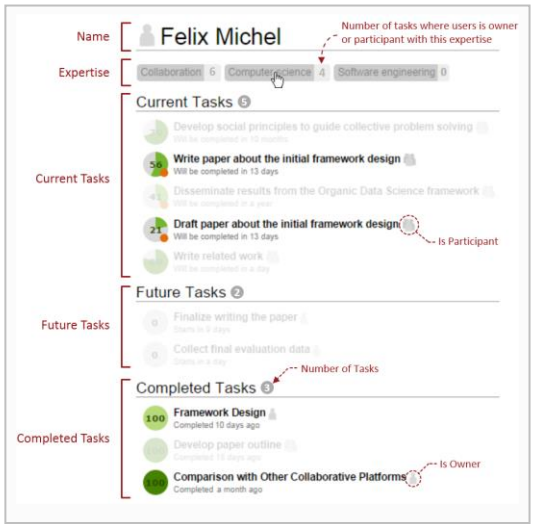
Tausczik, Y. R., Kittur, A., and Kraut, R. E. 2014. "Collaborative problem solving: a study of mathoverflow." *In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 355-367). ACM.

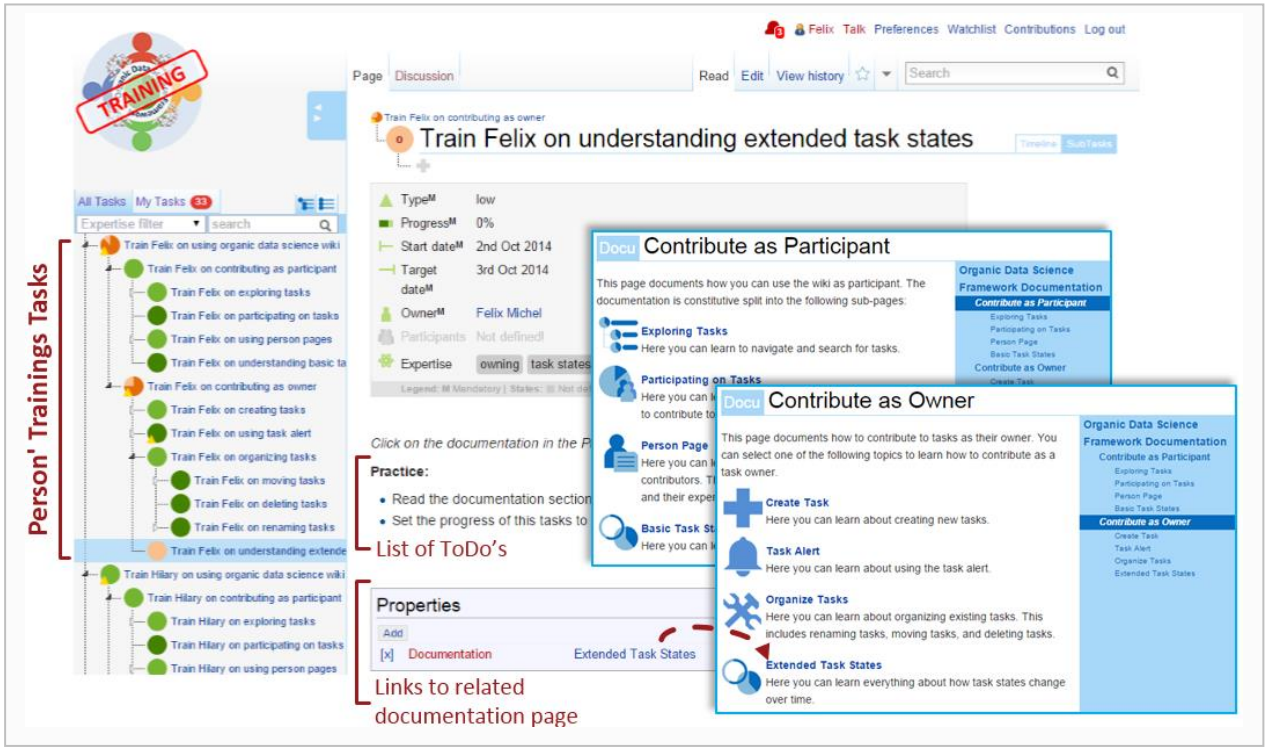
Van Merriënboer, J. J., Kirschner, P. A., and Kester, L. 2003. „Taking the load off a learner's mind: Instructional design for complex learning." *Educational psychologist*, 38(1), 5-13.

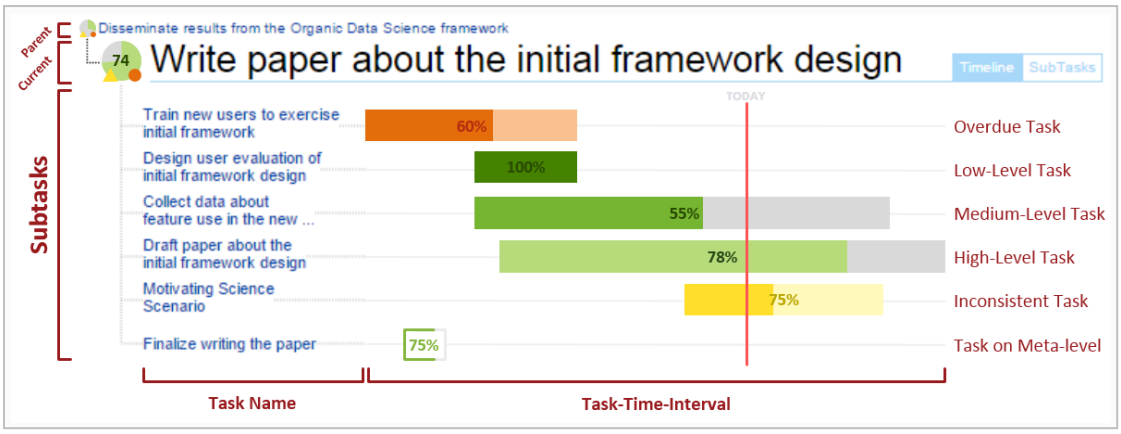
Appendix

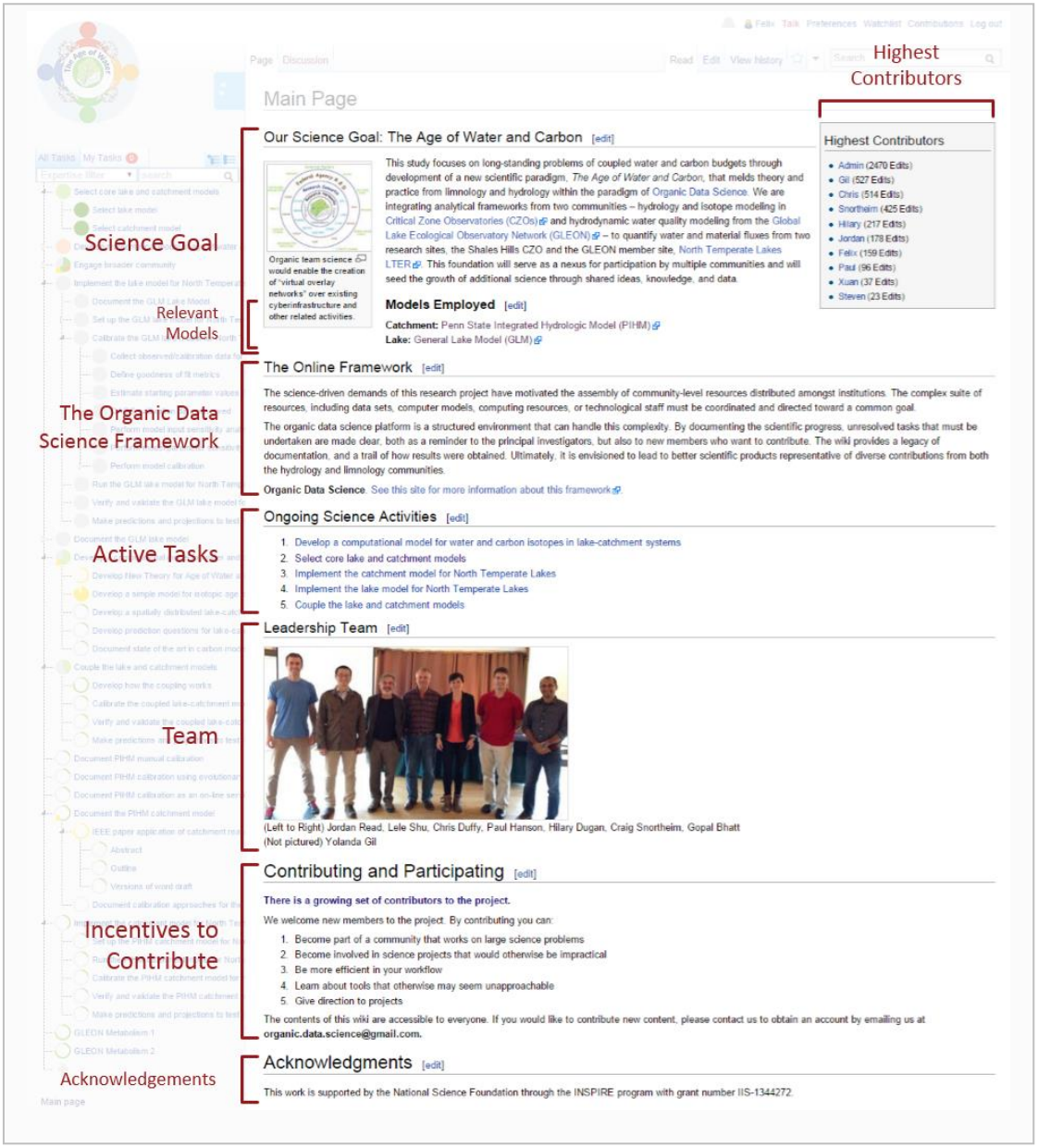
We provide screenshots of the Organic Data Science Wiki to illustrate our solution design.








Parent
Current
Disseminate results from the Organic Data Science framework
74
Write paper about the initial framework design
Timeline
SubTasks
TODAY
Subtasks
Train new users to exercise initial framework
60%
Overdue Task
Design user evaluation of initial framework design
100%
Low-Level Task
Collect data about feature use in the new ...
55%
Medium-Level Task
Draft paper about the initial framework design
78%
High-Level Task
Motivating Science Scenario
75%
Inconsistent Task
Finalize writing the paper
75%
Task on Meta-level
Task Name
Task-Time-Interval


Highest Contributors
Main Page
Our Science Goal: The Age of Water and Carbon [edit]
Highest Contributors
Admin (2470 Edits)
Gil (527 Edits)
Chris (514 Edits)
Snortheim (425 Edits)
Hilary (217 Edits)
Jordan (178 Edits)
Felix (159 Edits)
Paul (96 Edits)
Xuan (37 Edits)
Steven (23 Edits)
Science Goal
This study focuses on long-standing problems of coupled water and carbon budgets through development of a new scientific paradigm, The Age of Water and Carbon, that melds theory and practice from limnology and hydrology within the paradigm of Organic Data Science. We are integrating analytical frameworks from two communities – hydrology and isotope modelling in Critical Zone Observatories (CZOs) and hydrodynamic water quality modeling from the Global Lake Ecological Observatory Network (GLEON) – to quantify water and material fluxes from two research sites, the Shales Hills CZO and the GLEON member site, North Temperate Lakes LTER. This foundation will serve as a nexus for participation by multiple communities and will seed the growth of additional science through shared ideas, knowledge, and data.
Organic team science would enable the creation of "virtual overlay networks" over existing cyberinfrastructure and other related activities.
Relevant Models
Models Employed [edit]
Catchment: Penn State Integrated Hydrologic Model (PIHM)
Lake: General Lake Model (GLM)
The Online Framework [edit]
The Organic Data Science Framework
The science-driven demands of this research project have motivated the assembly of community-level resources distributed amongst institutions. The complex suite of resources, including data sets, computer models, computing resources, or technological staff must be coordinated and directed toward a common goal.
The organic data science platform is a structured environment that can handle this complexity. By documenting the scientific progress, unresolved tasks that must be undertaken are made clear, both as a reminder to the principal investigators, but also to new members who want to contribute. The wiki provides a legacy of documentation, and a trail of how results were obtained. Ultimately, it is envisioned to lead to better scientific products representative of diverse contributions from both the hydrology and limnology communities.
Organic Data Science. See this site for more information about this framework.
Ongoing Science Activities [edit]
Active Tasks
1. Develop a computational model for water and carbon isotopes in lake-catchment systems
2. Select core lake and catchment models
3. Implement the catchment model for North Temperate Lakes
4. Implement the lake model for North Temperate Lakes
5. Couple the lake and catchment models
Leadership Team [edit]
Team
(Left to Right) Jordan Read, Lele Shu, Chris Duffy, Paul Hanson, Hilary Dugan, Craig Snortheim, Gopal Bhatt
(Not pictured) Yolanda Gil
Contributing and Participating [edit]
Incentives to Contribute
There is a growing set of contributors to the project.
We welcome new members to the project. By contributing you can:
1. Become part of a community that works on large science problems
2. Become involved in science projects that would otherwise be impractical
3. Be more efficient in your workflow
4. Learn about tools that otherwise may seem unapproachable
5. Give direction to projects
The contents of this wiki are accessible to everyone. If you would like to contribute new content, please contact us to obtain an account by emailing us at organic.data.science@gmail.com.
Acknowledgments [edit]
Acknowledgements
This work is supported by the National Science Foundation through the INSPIRE program with grant number IIS-1344272.