Organic Data Science: Towards Task-based Online Communities for Open Collaboration in Science

Anonymized Authors
Affiliation
Address
e-mail address
Optional phone number

Anonymized Authors
Affiliation
Address
e-mail address
Optional phone number

Anonymized Authors
Affiliation
Address
e-mail address
Optional phone number

ABSTRACT
Although collaborative activities are paramount in science, little attention has been devoted to supporting on-line scientific collaborations. This paper presents an Organic Data Science framework to support scientific collaborations that revolve around complex science questions that require multi-disciplinary contributions to gather and analyze data, significant coordination to synthesize findings, and grow organically to accommodate new contributors as needed as the work evolves over time. The key idea is to open science by exposing science processes declaratively to enable broader participation through a task-based online community. We hope to make these kinds of scientific collaborations more common, reduce the coordination effort required, and lower the barriers to incorporating new collaborators through an intelligent user interface that supports open science processes.

Author Keywords
Collaboration interfaces, Organic Data Science.

ACM Classification Keywords
H.5.3. Information interfaces and presentation (e.g., HCI): Group and organization interfaces.

INTRODUCTION
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 [Ribes and Finholt 2009; Bos et al 2007]. 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 quest (e.g., the Human Genome Project). Our work focuses 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 knowledge, 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. Yet, these kinds of collaborations are needed in order to address major engineering and science challenges ahead (e.g., http://www.engineeringchallenges.org). Our goal is to develop a collaborative software platform that supports such scientific collaborations, and ultimately make them significantly more efficient and commonplace.

This paper presents an **Organic Data Science framework** to support scientific collaborations that revolve around complex science questions that require multi-disciplinary contributions to gather and analyze data, significant coordination to synthesize findings, and grow organically to accommodate new contributors as needed as the work evolves over time. The key idea is to open science by exposing science processes declaratively to enable broader participation. Science processes describe the what, who, when, and how of the activities pursued by the collaboration. The framework is still under development,

and it evolves to accommodate user feedback and to incorporate new collaboration features.

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.

There are a wide range of approaches that have been explored for collaboration, although they have not had much adoption in science practice. Collaborative user interfaces that have been used in science include semantic wikis (eg, [Huss et al 2010]), workflow repositories [De Roure et al 2009], and argumentation systems (e.g., [Introne et al 2013]). In addition, popular collaborative Web frameworks are also used in science, including code repositories, blogs, and wikis.

The paper begins with a motivating scenario of a complex science task that we are currently pursuing using this framework. We then introduce our task-centered approach for open scientific collaboration. We present our implemented framework, and a preliminary evaluation with user data collected to date.

MOTIVATING SCENARIO

The recognition that the future health of the world depends on provisioning of ecosystem services provided by fresh waters, including quantity and quality available for consumption, agriculture and aquaculture, industry, recreation, and carbon sequestration, has motivated an array of research and advocacy initiatives [MEA 2005; ILEC 2007; Levin and Clark 2009]. The resulting knowledge is represented in multiple disciplines, including hydrology [NRC: 2011, 2012], ecology [Foley et al. 2011], economics and security [Suweis et al. 2013].

Unfortunately, research programs dedicated to water are very fragmented. The Critical Zone Observatories (CZOs) focus on the interaction of water with soil, air, and living organisms on selected areas of the US. The Global Lake Ecological Observatory Network (GLEON) is a community focused on lake ecosystems. The Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) facilitates access to time series data about water. The Long Term Ecological Research (LTER) studies the ecosystem in particular sites spanning decades or centuries. There are many other organizations with relevant expertise. Despite great scientific advances and cross-connections among these collaborations, scientists still are challenged to quantify water and material fluxes that underpin aquatic ecosystems, and in some cases even understand the dominant mechanisms controlling them.

The scientific research that we are pursuing focuses on theoretical and experimental aspects of the isotopic "age" of water in watershed-lake systems. In this context, "age" is defined as the time since the water parcel and environmental tracer entered the system as precipitation. Both the hydrology and limnology communities have developed an observing system for isotope ratios of carbon, oxygen and hydrogen but with very different science questions. Our hypothesis is that the watershed-lake isoscape provides the experimental basis for predicting flow paths, residence times and the relative age of water in space and time, and that understanding these spatiotemporal patterns provides a deeper understanding of fundamental biogeochemical processes including carbon and nitrogen cycling within the lake catchment system. In order to pursue this research agenda, a collaboration needs to be formed that includes experts in models and data for lakes and for watersheds to develop a unified "isoscape" model for the watershed-lake system.

The aims of the domain research cut across disciplines and cross-institutional and geographic boundaries. An initial goal is a retrospective analysis based on a fully-coupled catchment-lake-groundwater hydrodynamic model parameterized from national data, calibrated with local data, and implemented to run climate and landuse change scenarios. As water age and the associated flowpaths are identified, scientists will use that information to infer the sources of organic carbon to lake-catchment ecosystems, their fluxes from the landscape to lakes, the fates as storage, conversion or export, and understanding of the uncertainties surrounding these quantities. Achieving these aims requires agreements and implementations of the data and model standards necessary for interoperability between hydrologic and ecological sciences, as well as a framework for integrating catchment-lake stable isotope analysis models.

A diverse and complex suite of resources, including data sets, computer models, computing resources, and researchers with diverse expertise must be coordinated and directed toward a common goal. The scientific goals of this project pose both technical and social challenges that require major amounts of coordination among participants. In addition, the science results will be stronger if more scientists join the collaboration, whether to contribute both data and analyses for additional sites or to contribute their expertise.

With current practices, this project would be very costly. The growth of the collaboration would take years to reach critical mass. Significant effort would have to be devoted to coordinating activities.

Our goal is to develop a collaboration framework that helps scientists to harmonize the resources needed for the project, to establish the processes for carrying out research within the project, and to foster the growth of the collaboration. This new framework must make the collaboration more efficient in terms of time and cost, and must be able to attract newcomers.

APPROACH

The key features of our approach are:

1. **providing a task-oriented nexus driven by science goals** that connects scientists together, organizing tasks to help scientists track where they can contribute and when, as well as their past contributions
2. **incorporating principles from social sciences research on successful on-line collaborations**, including best practices for retention and growth of the community
3. **opening the science process to expose all tasks and activities publicly**, so all participants (especially newcomers) can immediately see the work being done and the tasks they can contribute to

Task-Centered Collaborative Spaces

In practice, the contributors to the organic data science framework form an organization. We use tasks as an organizational mechanism for coordination. Task organization and processes have been shown to be a key aspect of collaboration in science laboratories [Chandrasekaran and Nernessian 2015]. 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.

[Polanyi 1983] 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]. In our project, we aim at externalizing the tacit knowledge of researchers to formulate and resolve tasks collaboration. While we focus on science processes in this paper, processes and tasks have been found o be important for the productivity of knowledge workers in an organizational context [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]. Temporal aspects of task achievement are also important. In project management, the duration estimates and resource selection have been found to be important [Pietras and Coury 94]. The user interface should be designed so users have some initial structure to express tasks. [Van Merrienboer 97] proposes the use of process worksheets to guide students through complex tasks. [Mahling and Croft 88] also found that the formulation of tasks is greatly improved through form-based interfaces.

Social Principles

There are numerous studies about successful on-line communities [Kraut and Resnick 2011]. Many studies are focused on Wikipedia and other wiki-style frameworks, 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; Leskovec et al 2010], critical mass of contributors [Raban et al 2010], coordination across contributions [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]. These studies suggest a number of principles for the design of our on-line collaboration framework.

Figure 1 summarizes the social principles that we are using in our approach. We follow the organization used in [Kraut and Resnick 2011], but we focus here 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). The principles are written to be self-explanatory, and in the next section we will explain how they map to features in our user interface (marked with numbers at the end of each line).

Opening Science Process

We find inspiration in the Polymath project, set up to collaboratively develop proofs for mathematical theorems [Nielsen 2011; Gowers 2009a], where professional mathematicians collaborate with volunteers that range from high-school teachers to engineers to solve mathematics conjectures. The collaboration is centered around tasks, that contributors create, decompose, reformulate, and resolve. This project uses common Web infrastructure for collaboration, interlinking public blogs for publishing problems and associated discussion threads [Nielsen 2013] with wiki pages that are used for write-ups of basic definitions, proof steps, and overall final publication [Gowers 2013]. Interactions among contributors to share tasks and discuss ideas are regulated by a simple set of guidelines that serve as social norms for the collaboration [Gowers 2009b]. The growth of the community is driven by the tasks that are posted, as tasks are decomposed into small enough chunks that potential contributors can see a way to contribute.

Another project that has exposed best practices of a large collaboration is ENCODE [Birney 2012; Nature 2012]. 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.

1. **Starting communities**
   - 1.1. Carve a niche of interest, scoped in terms of topics, members, activities, and purpose ⓪
   - 1.2. Relate to competing sites, integrate content ⓪
   - 1.3. Organize content, people, and activities into subspaces once there is enough activity ⓪①⑧
   - 1.4. Highlight more active tasks ①②④
   - 1.5. Inactive tasks should have "expected active times" ②⑥
   - 1.6. Create mechanisms to match people to activities ①②
2. **Encouraging contributions through motivation**
   - 2.1. Make it easy to see and track needed contributions ①②③④⑤⑥⑦⑨⑩
   - 2.2. Ask specific people on tasks of interest to them ②⑨
   - 2.3. Simple tasks with challenging goals are easier to comply with ①⑧
   - 2.4. Specify deadlines for tasks, while leaving people in control ②③④⑦
   - 2.5. Give frequent feedback specific to the goals ②④⑥⑨⑩
   - 2.6. Requests coming from leaders lead to more contributions ②
   - 2.7. Stress benefits of contribution ⓪
   - 2.8. Give (small, intangible) rewards tied to performance (not just for signing up) ⑨
   - 2.9. Publicize that others have complied with requests ④
   - 2.10. People are more willing to contribute: 1) when group is small, 2) when committed to the group, 3) when their contributions are unique ①③④⑧⑨
3. **Encouraging commitment**
   - 3.1. Cluster members to help them identify with the community ③②⑨
   - 3.2. Give subgroups a name and a tagline ①②③
   - 3.3. Put subgroups in the context of a larger group ①③④
   - 3.4. Make community goals and purpose explicit ①③
   - 3.5. Interdependent tasks increase commitment and reduce conflict ①②③⑤⑨
4. **Dealing with newcomers**
   - 4.1. Members recruiting colleagues is most effective ⓪
   - 4.2. Appoint people responsible for immediate friendly interactions ⑪
   - 4.3. Introducing newcomers to members increases interactions ⑪
   - 4.4. Entry barriers for newcomers help screen for commitment ⑪
   - 4.5. When small, acknowledge each new member ⓪
   - 4.6. Advertise members particularly community leaders, include pictures ⓪
   - 4.7. Provide concrete incentives to early members ⓪
   - 4.8. Design common learning experiences for newcomers ⑪
   - 4.9. Design clear sequence of stages to newcomers ⑪
   - 4.10. Newcomers go through experiences to learn community rules ⑪
   - 4.11. Provide sandboxes for newcomers while they are learning ⑪
   - 4.12. Progressive access controls reduce harm while learning ⑪

Figure 1. Selected social principles from [Kraut and Resnick 2011] 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). The circled numbers at the end of each line indicate user interface features that implement these principles, illustrated in Figures 3, 4, and 8.

5. **Best practices from Polymath**
   - 5.1. Permanent URLs for posts and comments, so others can refer to them ①
   - 5.2. Appoint a volunteer to summarize periodically ⓪
   - 5.3. Appoint a volunteer to answer questions from newcomers ⑪
   - 5.4. Low barrier of entry: make it VERY easy to comment ⑪
   - 5.5. Advance notice of tasks that are anticipated ②⑥⑩
   - 5.6. Keep few tasks active at any given time, helps focus ⓪
6. **Lessons learned from ENCODE**
   - 6.1. 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 ⓪
   - 6.2. Written and publicly accessible rules to transfer work between groups, to assign credit when papers are published, to present the work ⓪
   - 6.3. Quality inspection with visibility into intermediate steps ①②③⑤⑥⑧⑩
   - 6.4. Export of data and results, integration with existing standards ⓪

Figure 2. Selected best practices from the Polymath [Nielsen 2011] project and lessons learned from ENCODE [Nature 2012] that can be applied to the initial design of our Organic Data Science framework. The circled numbers at the end of each line indicate user interface features that implement these principles, illustrated in Figures 3, 4, and 8.

FRAMEWORK

In this section, we describe our current implementation of the Organic Data Science framework. It is built as an extension of the Semantic Media Wiki platform [Krötzsch et al 2011; Bry et al 2012], and uses its semantic capabilities to structure the content of the site, including task properties, user properties. The semantic wiki provides an intuitive user interface that hides from users any formal semantic notation [Gil 2013; Bry et al 2012]. The site is accessible from http://www.organicdatascience.org.

We highlight here the major features of the user interface in terms of the Social Design Principles in Figure 1 and the best practices in Figure 2, showing where they appear through screenshots of various pages.

0 Welcome Page: Figure 3 shows the main page of the site, which is set up to describe clearly the science and technical objectives of the project, to display a summary of currently active tasks, and to show the leadership and major contributors. 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. The model and contributor lists are dynamically generated from the current content with a semantic wiki query, so they are always up to date. Anyone can see the contents of the site, so 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 (see feature 11).

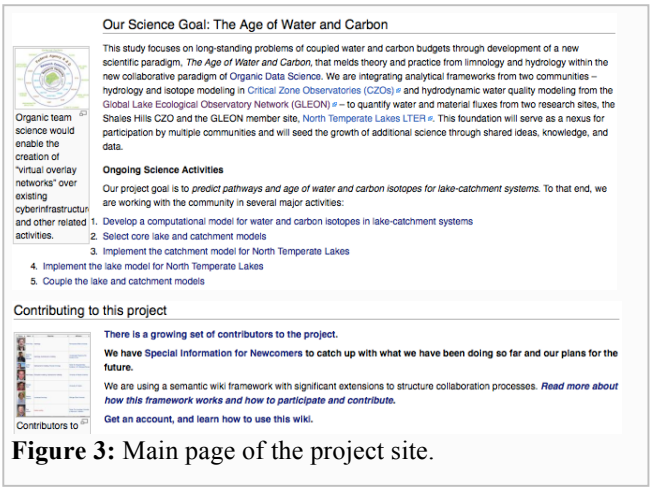

Figure 3: Main page of the project site.

1 Task Representation: 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. The name of the task is shown at the top, with its parent task above it in much smaller font. An icon showing the status of the task is shown next to the name (feature 10). Below that, the user is shown either a subtask navigation or a timeline (features 5 and 6). After that, a gray box shows all the metadata (feature 2). Everything below this box is page content, i.e., text that describes the actual work involved in doing a task

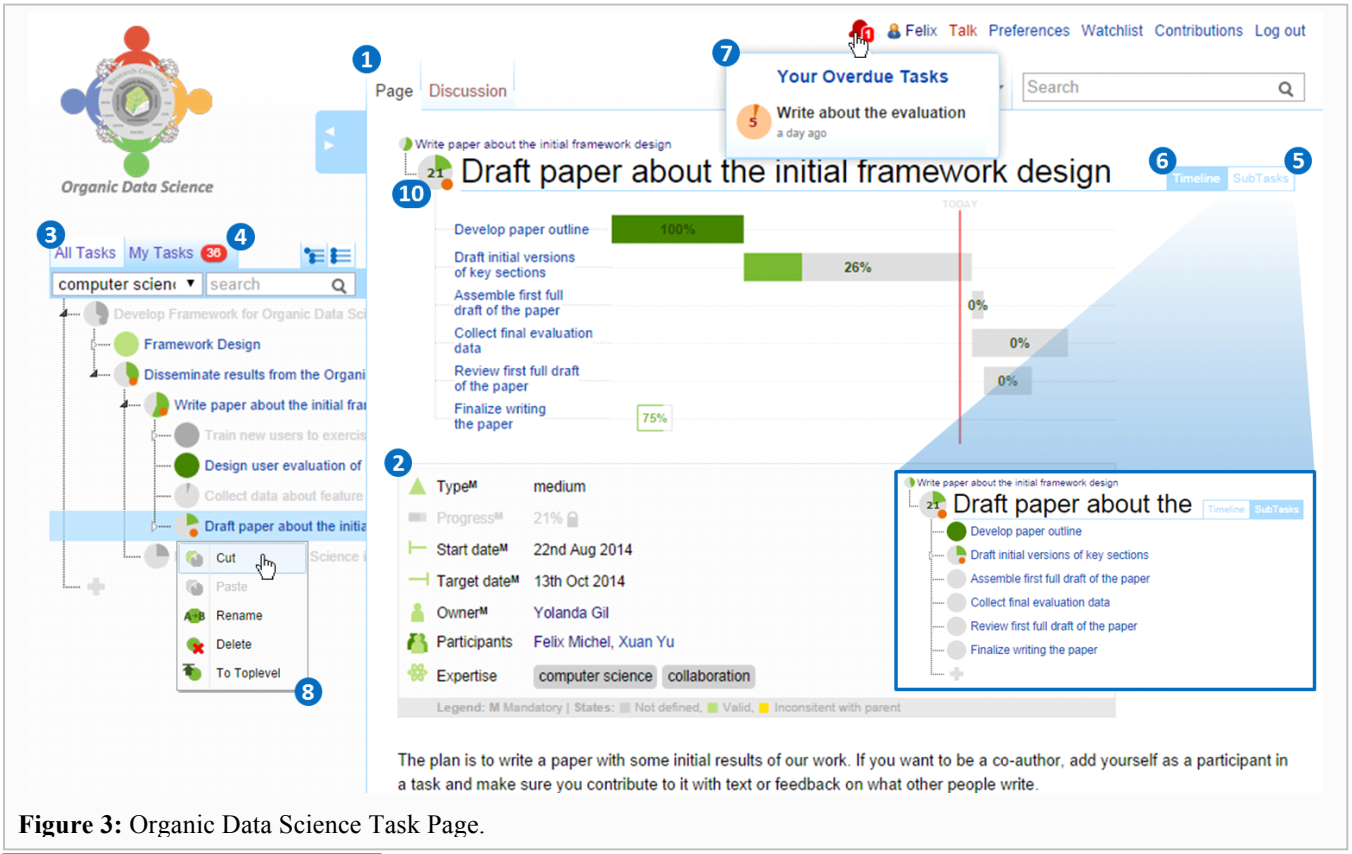

Figure 3: Organic Data Science Task Page.

as well as the results when it is completed.

❷**Task Metadata:** Task metadata are major properties of the task, such as begin and end times, which enable the system to assist users to manage tasks. All task metadata is stored in the wiki as semantic properties of the task page. We distinguish between several categories of metadata. *Pre-defined metadata* are properties of tasks that the system will use to assist users to manage tasks (features 4, 7, 9, 10). 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), a 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. *Dynamically-defined metadata* (not shown in Figure 3 allow users to create new properties on the fly that help group tasks with domain-specific features, for example tasks that have to do with calibration of models or tasks that are outreach tasks. An important required metadata property is the task type, which is provided by the user and helps the system estimate the progress and status of tasks. High-level tasks are assumed to have a high abstraction grade and a 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 green colors in the task icon, with high-level task 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 higher-level tasks is calculated dependent on the task type and the start and 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. 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 are green, and properties that are inconsistent with properties of the parent task are yellow.

❸**Task Navigation:** Similar to the well-known hierarchical folder navigation used in operating systems, we provide hierarchical task navigation. The user can expand the nested task structure until a leaf task is reached. The user can search tasks based on words used in the task titles, and can select a task expertise as a filter for the search. Tasks which do not match with the filter are hidden, except for the parent tasks of the subtasks matching the filter are shown as a way to provide context but faded out.

❹**Personal Worklist:** The worklist contains a subset of tasks from the task explorer, and they are the tasks which contain the user as owner or as participant. A red counter indicates the current number of tasks in the user's worklist.

❺**Subtask Navigation:** Subtasks of the currently opened task are presented as part of the task page. The navigation works similar to the Task Navigation. No filter and search option is provided in this navigation.

❻**Timeline Navigation:** All subtasks are represented based on their start, end times, and completion status in a visualization based on a Gantt chart. Start and target date define the position. Tasks with completed required metadata are shown as a gray rectangle, with the percentage of completed work in green. Tasks with incomplete required metadata are represented with an empty rectangle and placed early on the chart. The timeline shows overdue tasks with orange bars and inconsistent tasks with a yellow bar.

❼**Task Alert:** A task alert occurs when a task is not completed and the target date is passed. Only the task owner gets this alert notification. A red alert bell with a small number indicates the number of overdue tasks. The owner is responsible for resolving this by completing the task, getting other users to complete the task, or delaying the target date if appropriate.

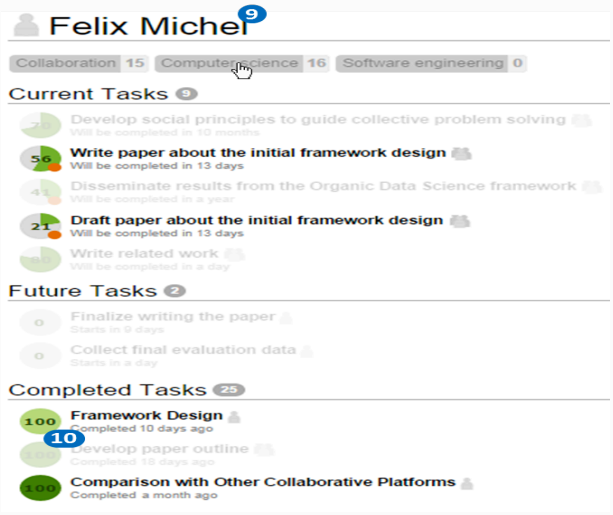

Figure 4: Organic Data Science Person Page.

❽**Task Management:** The interface supports actions like creating, renaming, moving and deleting tasks. For usability reasons, all these actions can be reversed. Subtasks can be created in the Task Navigation or the Subtask Navigation by using the plus button below the last task.

Root tasks can only be crated in the Task Navigation. If the Task Explorer has an expertise selected or/and the "my task" tab is used, the new task will have that metadata. Renaming, moving, and deleting tasks is done from the Task Navigation Deleting a task implies deleting all its subtasks. Tasks can be moved with the cut and paste operations. Moving a task to the root level works only with the "To top level" action. All moving actions can cause inconsistencies in the temporal state of tasks, for example if the time interval of the moved task does not fit into the time interval of its new parent task. The same problem can occur with the task type, for example when a high-level task is moved as subtask under a lower level. Task states that are inconsistent are highlighted in yellow (feature 10), and parent tasks indicate inconsistency in their subtasks with a small yellow triangle.

❾**User Tasks and Expertise:** The interface allows users to easily see what others are planning to work on or have worked on in the past. This creates a transparent work process. It also makes it easy for newcomers to browse the tasks of other users that share their expertise and find tasks of interest as well as ongoing tasks where they could get involved. The top of every user page contains a user icon followed by the user name. Next, the interface shows the user's expertise metadata property values. Hovering over a certain expertise value fades out all tasks that are not associated with that expertise.

❿**Task State:** The state of every task is summarized with a task icon next to the task name. Figure 5 illustrates how the system uses the task metadata to generate the task state. The left 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 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 (shown in the Task Explorer in Figure 3). Figure 6 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 did 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 notice 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 in the Timeline Navigation (feature 6). Figure 7 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 late. 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 and in the user pages.

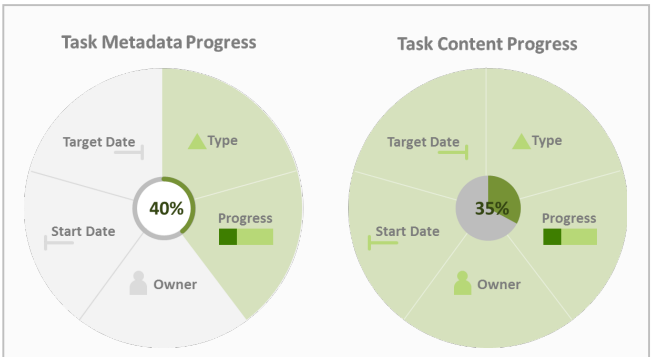

Figure 5: Conceptual task state estimation.

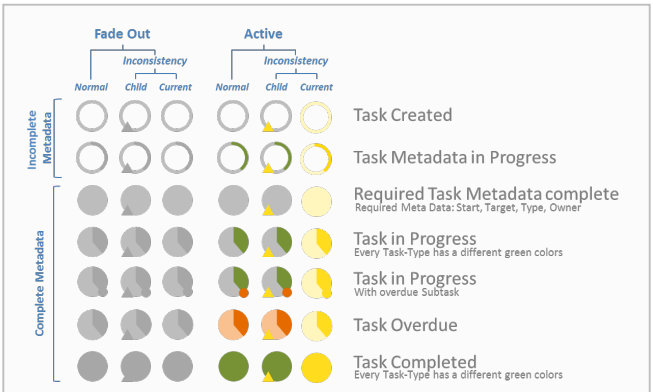

Figure 6: Possible task states.

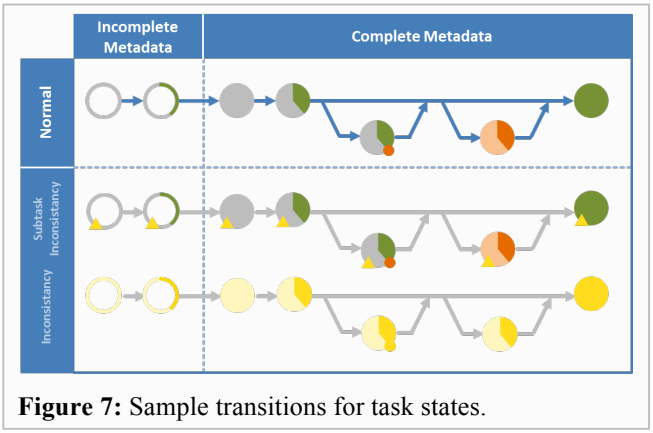

Figure 7: Sample transitions for task states.

⓫**Training new members**: We set up separate site[2] to train new users. This training site also uses the Organic Data Science framework, so it has the same features presented above. A new user is given a set of predefined training

[2] http://skc.isi.edu/smw/ods_training/

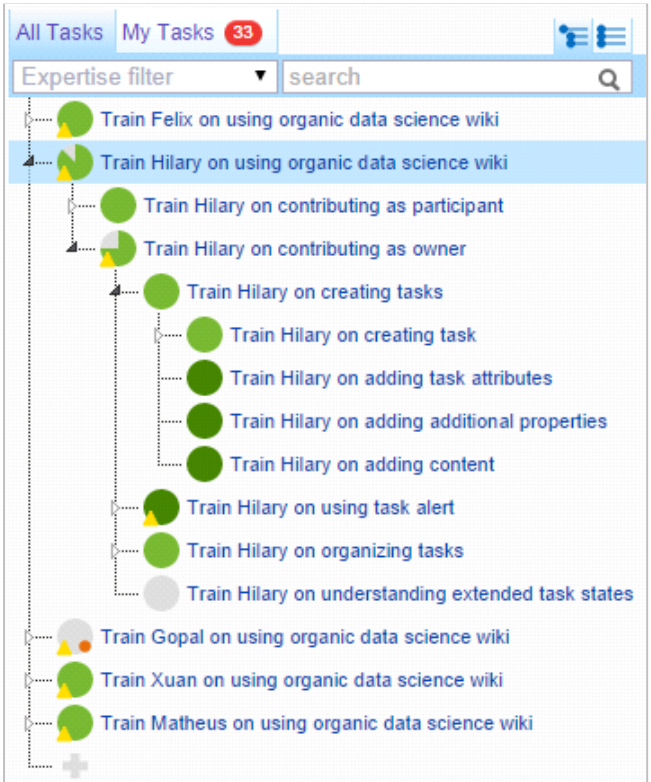

Figure 8: Tasks in the training site for a new member.

tasks, shown in Figure 8, each for learning and practicing a different feature of the interface. The training tasks follow the structure of the documentation pages, and allow new users to practice by using the same interface as they will use in the main site. As they complete the tasks, users can see the task status changing. The training is divided in two phases. The first phase trains them to contribute to existing tasks. The second phase trains them to create new tasks and to manage them as owners. One person in the collaboration is always assigned to help new users with their training, and is available by email to answer questions. This appointment rotates as new members become more experienced and can contribute in this capacity.

In summary, the system is designed to: 1) help users by organizing the collaboration around tasks, so that task contributors and progress can be easily tracked and highlighted, 2) manage contributors using social design principles and best practices from prior research in on-line communities, and 3) expose the scientific research process by making all the information about tasks open and widely accessible on the Web.

We have described here the current implementation of the system. The system continues to be extended based on feedback from the contributors. At the moment, this is done through email requests, but we are developing a more formal mechanism to create subtasks of the main project task to "Develop the Organic Data Science Framework." In addition, we have instrumented the system to record the usage of the site, so we can track what features work well and which ones are not popular. As the community grows, additional social design principles will be incorporated into the framework as new features.

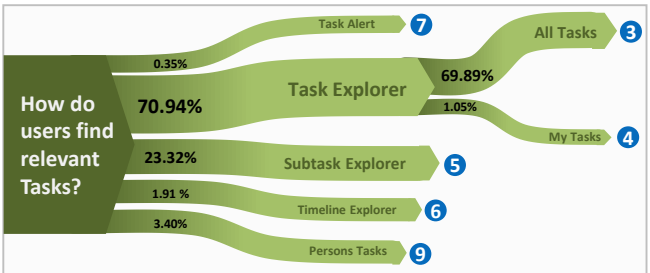

Figure 9: Finding Tasks via Features.

EVALUATION

We present an evaluation of our current implementation of the Organic Data Science approach. The site has been active since January 2014 and has been in use as new features were rolled out. It currently has 18 registered users, and contains 122 tasks. All task pages together have been accessed more than 2,900 times to date. All person pages together have been accessed 328 times. We instrumented our framework once all the features described above were rolled out. Within 10 weeks we collected around 19,000 log entities about how each user interface feature was being used. Additionally we organized and write this paper collaboratively with our organic data science framework.

Is the Framework Helping Users Manage Tasks?

We evaluated this based on the logs for the features that help users manage tasks.

How do users find relevant tasks? Figure 9 shows what features are used by users to open task pages. Most users used the Task Navigation feature to find task pages. This is probably because this feature gives users an overview over all tasks, drill down quickly, and apply specific filters. The Task Alert feature was not used very often, but we expect that this feature will be more important as the group faces deadlines (such as the writing of this paper, an upcoming scientific workshop, etc.).

What features are used to manage tasks? Figure 10 shows heat maps that illustrate in red where users click most. Every heat map represents the clicks on one single page. On the left is the main page of the site, most clicks are on the Task Navigation feature. On the right is a task page, showing that most clicks occur in areas where many of our task features are situated.

Is the Framework Helping Users to Collaborate?

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 these are tasks that were recently created and did not even have an owner.

Figure 10: Heat maps for two pages showing user clicks.

We did not filter out data for tasks that were renamed or deleted. All results are illustrated in Figure 11.

Ⓐ*How many tasks are viewed by more than one person?* Figure 11(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 be lower as the project progresses, particularly as it gets closer to completion.

Ⓑ*How many tasks have more than one person signed up?* Figure 11(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.

Ⓒ*How many tasks have more than one person editing task metadata?* Figure 11(c) shows these results. Currently 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 tat non-owners have taken an interest in the management of the tasks.

Ⓓ*How many tasks have more than one person editing their content?* This is shown in Figure 11(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.

What does the social network of collaborators look like? 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 with the strength of edges. The result is illustrated in Figure 12. 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. The collaboration group developing the software for the framework is at the bottom, while the collaboration group working on the science goals of

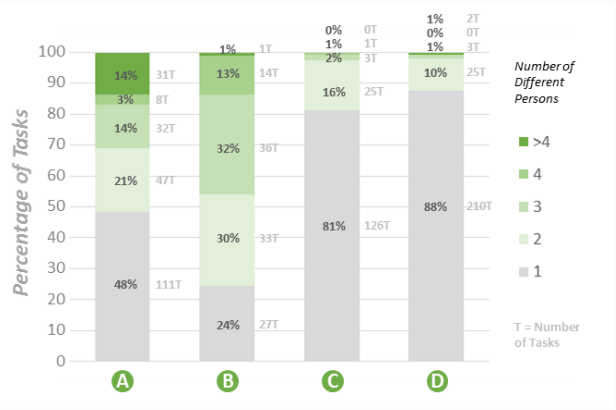

Figure 11: Task Collaboration Evaluation.

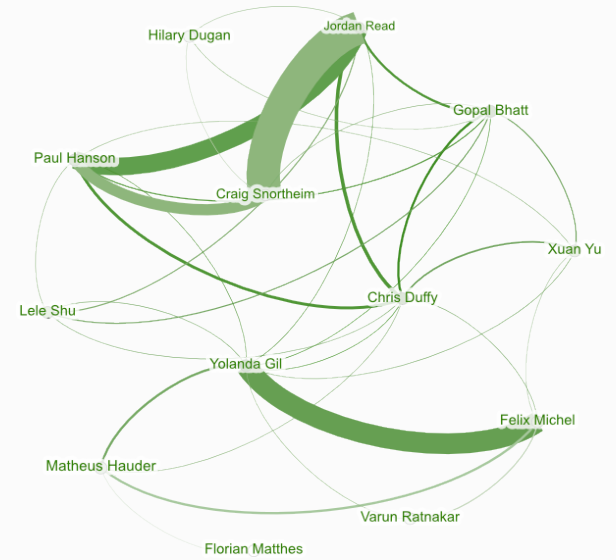

Figure 12: Organic Data Science Collaboration Graph.

studying the age of water is at the top. There are many links across these groups, as both are concerned with the design of the overall approach to Organic Data Science.

Is the Framework Helping Newcomers?

To evaluate how valuable the documentation in combination with the training is, we evaluate:

How often is the documentation accessed after training? The log did not contain any data showing access to documentation after people have finished the training. We asked the new users to confirm this, which they did.

How often are tasks deleted shortly after creation by a given user? This question may help us understand if new users make mistakes in creating tasks. One user deleted a total of 3 tasks within five minutes after creation. Interestingly, this happened while they were practicing in the training site. Other users did not delete any tasks.

What is the new user's total training time? Newcomers who finished our training estimated it took around one hour total to train as task contributors and task owners.

RELATED WORK

We discuss related work in scientific collaboration, collaboration systems, and task-centered user interfaces.

Scientific Collaboration

[Bos et al 2007] did a comprehensive study of scientific collaborations and propose seven types: *1) Shared Instruments*, where instruments or sensors are used by a community (e.g., National Ecological Observatory Network [NEON 2014]); *2) Community Data Systems*, where a data resource is maintained and used by a community (e.g., the Protein Data Bank [Berman et al 2000]), *3) Open Community Contribution Systems*, where tasks are carried out by a community including citizen scientists (e.g., the GalaxyZoo project for labeling galaxy images [Lintott 2010]), *4) Virtual Communities of Practice*, where a community shares interest in specific research topics (e.g., the Global Lake Ecological Observatory Network [GLEON 2014]), *5) Virtual Learning Communities*, where the purpose is to learn through the collaboration (e.g., the VIVO research network [Krafft et al 2010]), *6) Distributed Research Centers*, where several institutions collaborate in a funded project (e.g., the ENCODE genomics project [Nature 2012], and *7) Community Infrastructure Projects*, where a community shares computing and software infrastructure (e.g., the Community Surface Dynamics Modeling System [Peckham et al 2013]). Our work has some of the properties of a distributed research center (6), and is an open community contribution system (3) but without the prescribed tasks typically found there. Organic Data Science can be considered a new type of collaboratory, where the tasks are defined on the fly as the project progresses and the collaboration includes unanticipated contributors.

[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. 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.

On-Line Collaboration Systems

THIS PARAGRAPH NEEDS WORK. Several different approaches help to manage a task structure. Rhythms help to establish structures but plans should be flexible enough to react on changes [Steinhardt and Jackson. 2014]. Visual feedback helps to increase the task resumption rate with less stress [Liu et al 2014]. Studies on different tools illustrate needed improvements for collaboration. A study of MathOverlow shows how the quality of answers can be improved collaboratively [Tausczik et al 2014]. Another study on Electronic Lab Notebooks shows the need of improving structuring knowledge in an ad-hoc and simple manner [Oleksik et al 2014]. Simple management mechanisms help to enforce collaboration. An analysis of Wikipedia shows a continuously increasing readership and a decreasing contribution since 2007 and the resulting need of a task-centered contribution organization [Morgan et al 2014]. A communication board is needed to create the opportunity of a shared decision making, illustrated on patient data [Kane et al 2013].

Argumentation interfaces facilitate the collaborative synthesis of diverse ideas [Buckingham-Shum 2006], and have been used in the context of science. [Introne et al 2013] describe the Climate CoLab, a collaborative environment for climate research. It offers argumentation structures, where evidence and hypotheses from different scientists can be compared and integrated to create a common view on climate research. This work, however, does not focus on supporting science research tasks while they are being carried out, only on organizing results of scientific work done elsewhere.

Task-Centered User Interfaces

Some task-oriented collaboration systems have been developed for information seeking tasks (e.g., Web search). An example is Kolline [Filho et al 2010], which supports the collaboration is between inexperienced users that need help from more advanced users. Our goal is to support tasks that have interrelated subtasks and that involve collaboration among peers.

Other work on managing tasks in on-line environments addresses tasks for remote workers, such as microtasks in Amazon Mechanical Turk [Park et al 2014; Kamar et al 2012]. The workers are not explicitly coordinating the work, and the tasks are pre-defined for them and tend to be repetitive across workers.

User tasks are sometimes inferred from their use of the interface [Steichen et al 2013]. These tasks concern interface use, rather than coordination.

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; Gil et al 2011]. 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.

CONCLUSION

We have presented a novel on-line collaboration framework to support organic data science. 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 of the project. We have specific hypotheses about how the maturity of the project will affect the management of tasks, about how the growth of the community 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 community in quantitative terms.

MAX PAGES IS 10, NOT COUNTING BIBLIOGRAPHY.

REFERENCES

[Auer et al 06] Soren Auer, Sebestian Dietzold, and Thomas Riechert. "OntoWiki - A Tool for Social, Semantic Collaboration." 5th International Semantic Web Conference, 2006.

[Auer et al 07] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, Zachary Ives. "DBpedia: a nucleus for a web of open data." Proceedings of the 6th international semantic web conference, 2007.

[Berman et al 2000] "The Protein Data Bank." H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. Nucleic Acids Research, 28: 235-242, 2000.

[Birney 2013] "Lessons for big data projects." Ewan Birney. Nature, Special Issue on the ENCODE project, 6 September 2012.

[Bos et al 2007] "From Shared Databases to Communities of Practice: A Taxonomy of Collaboratories." Nathan Bos, Ann Zimmerman, Judith S. Olson, Jude Yew, Jason Yerkie, Erik Dahl, Gary M. Olson. Journal of Computer-Mediated Communication 12(2): 652-672 (2007).

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

[Bry et al 2012] François Bry, Sebastian Schaffert, Denny Vrandečić and Klara Weiand. "Semantic Wikis: Approaches, Applications, and Perspectives." Lecture Notes in Computer Science, Reasoning Web. Semantic Technologies for Advanced Query Answering, Volume 7487, 2012.

[Buckingham-Shum 2006] "Hypermedia Support for Argumentation-Based Rationale: 15 Years on from gIBIS and QOC." Simon J. Buckingham Shum, Albert M. Selvin, Maarten Sierhuis, Jeffrey Conklin, Charles B. Haley, Bashar Nuseibeh. In "Rationale Management in Software Engineering", Allen H. Dutoit, Raymond McCall, Ivan Mistrik, and Barbara Paech (Eds). Springer-Verlag, 2006.

[Chandrasekaran and Nersessian 2015] Building Cognition: The Construction of Computational Representations for Scientific Discovery. Chandrasekharan, S. & Nersessian, N.J. To appear in Cognitive Science, 2015. Available from http://www.cc.gatech.edu/aimosaic/faculty/nersessian/papers/Building-cognition_Cogsci_Final_manuscript.pdf

[Chin et al 2002] "New paradigms in problem solving environments for scientific computing." George Chin, Jr., L. Ruby Leung, Karen Schuchardt, Debbie Gracio. IUI 2002.

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

[De Roure et al 2009] De Roure, D.; Goble, C.; Stevens, R, The design and realization of the myexperiment virtual research environment for social sharing of workflows, Future Generation Computer Systems 25 (5), 2009.

[Filho et al 2010] "Kolline: a task-oriented system for collaborative information seeking." Fernando Marques Figueira Filho, Gary M. Olson, Paulo Lício de Geus. SIGDOC, 2010.

[Gil 2013] "Social Knowledge Collection." Gil, Y. In Handbook of Human Computation, P. Michelucci (Ed). Springer, 2013.

[Gil et al 2011] "Wings: Intelligent Workflow-Based Design of Computational Experiments." Gil, Y.; Ratnakar, V.; Kim, J.; Gonzalez-Calero, P. A.; Groth, P.; Moody, J.; and Deelman, E. IEEE Intelligent Systems, 26(1). 2011.

[Gil and Ratnakar 2013] "Knowledge Capture in the Wild: A Perspective from Semantic Wiki Communities." Gil, Y.; and Ratnakar, V. In Seventh ACM International Conference on Knowledge Capture (K-CAP), Banff, Canada, 2013.

[GLEON 2014] GLEON. The Global lakes Ecological Observatory Network. From http://www.gleon.org. Last accessed October 7, 2014.

[Huss et al 2010] "The Gene Wiki: community intelligence applied to human gene annotation." Huss JW, 3rd, Lindenbaum P, Martone M, Roberts D, Pizarro A, Valafar F, Hogenesch JB and Su AI. Nucleic Acids Res., 38:D633-D639, 2010.

[Hutchins 1995] "How a Cockpit remembers its speeds", E. Hutchins, Cognitive Science, 19, 1995.

[Introne et al 2013] "Solving Wicked Social Problems with Socio-computational Systems." Joshua Introne, Robert Laubacher, Gary M. Olson, Thomas W. Malone. KI - Künstliche Intelligenz 27(1): 45-52 (2013).

[Kamar et al 2012] “Combining human and machine intelligence in large-scale crowdsourcing.” Ece Kamar, Severin Hacker, Eric Horvitz. AAMAS 2012.

[Kane et al 2013] "Shared Decision Making Needs a Communication Record." Bridget Kane, Pieter Toussaint and Saturnino Luz. Computer Supported Cooperative Work and Social Computing (CSCW), San Antonio, Texas, February 2013.

[Kittur et al 08] Aniket Kittur, Bongwon Suh, Ed H. Chi. “Can you ever trust a wiki? Impacting perceived trustworthiness in Wikipedia.” Proceedings of the ACM conference on Computer supported cooperative work, 2008.

[Kittur and Kraut 08] Aniket Kittur, Robert E. Kraut. “Harnessing the wisdom of crowds in Wikipedia: Quality through coordination.” Proceedings of the ACM conference on Computer supported cooperative work, 2008.

[Kittur et al 09] Aniket Kittur, Bryant Lee, Robert E. Kraut. “Coordination in collective intelligence: the role of team structure and task interdependence.” Proceedings of the 27th international conference on Human factors in computing systems, 2009.

[Kittur and Kraut 10] Aniket Kittur, Robert E. Kraut. “Beyond Wikipedia: coordination and conflict in online production groups.” Proceedings of the 2010 ACM conference on Computer supported cooperative work.

[Krafft et al 2010] “VIVO: enabling national networking of scientists.” D Krafft, N Cappadona, B Caruso, J Corson-Rikert, M Devare, B Lowe, and VIVO Collaboration. Conference on Web Science (WebSci), Raleigh, NC, April 2010.

[Kraut and Resnick 2011] “Building Successful Online Communities: Evidence-Based Social Design.” Robert E. Kraut and Paul Resnick. MIT Press, 2011.

[Krötzsch et al 07] Markus Krötzsch, Denny Vrandecic, Max Völkel, Heiko Haller, Rudi Studer. “Semantic Wikipedia.” Journal of Web Semantics, 5(4), pages 251-261, December 2007.

[Krötzsch et al 2011] Markus Krötzsch, Denny Vrandecic: Semantic MediaWiki. Foundations for the Web of Information and Services 2011: 311-326

[Kuhn 09] Tobias Kuhn. “AceWiki: A Natural and Expressive Semantic Wiki.” Proceedings of the Fifth International Workshop on Semantic Web User Interaction (SWUI 2008), CEUR Workshop Proceedings, Volume 543, 2009.

[Lam et al 10] Shyong (Tony) K. Lam, Jawed Karim, John Riedl. “The effects of group composition on decision quality in a social production community.” Proceedings of the 16th ACM international conference on supporting group work, 2010.

[Leskovec et al 10] J. Leskovec, D. Huttenlocher, J. Kleinberg. Governance in Social Media: A case study of the Wikipedia promotion process. Proceedings of the AAAI International Conference on Weblogs and Social Media (ICWSM), 2010.

[Lintott et al 10] Chris Lintott, Kevin Schawinski, Steven Bamford, An"e Slosar, Kate Land, Daniel Thomas, Edd Edmondson, Karen Masters, Robert C. Nichol, M. Jordan Raddick, Alex Szalay, Dan Andreescu, Phil Murray, Jan Vandenberg. “Galaxy Zoo 1: data release of morphological classifications for nearly 900,000 galaxies”. Monthly Notices of the Royal Astronomical Society, 2010

[Liu et al 2014] "Supporting Task Resumption Using Visual Feedback." Yikun Liu, Yuan Jia, Wei Pan and Mark Pfaff. Computer Supported Cooperative Work and Social Computing (CSCW), Baltimore, Maryland, February 2014.

[Mahling and Croft 1988]. “Relating human knowledge of tasks to the requirements of plan libraries.” D. E. Mahling and W. B. Croft. International Journal of Human-Computer Studies, Vol. 31, 1988.

[Mahling and Croft 1993]. “Acquisition and Support of Goal-Based Tasks.” D. E. Mahling and W. B. Croft. Knowledge Acquisition, Vol. 5, 1993.

[Mao et al 2013] “Volunteering Versus Work for Pay: Incentives and Tradeoffs in Crowdsourcing.” Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan E. Schwamb, Chris J. Lintott, Arfon M. Smith. HCOMP 2013.

[McGuinness et al 2006] Deborah L. McGuinness, Honglei Zeng, Paulo Pinheiro da Silva, Li Ding, Dhyanesh Narayanan, and Mayukh Bhaowal. “Investigations into Trust for Collaborative Information Repositories: A Wikipedia Case Study.” Proceedings of the Workshop on Models of Trust for the Web, 2006.

[Morgan et al 2014] "Editing Beyond Articles: Diversity & Dynamics of Teamwork in Open Collaborations." Jonathan T. Morgan, Michael Gilbert, David W. McDonald and Mark Zachry. Computer Supported Cooperative Work and Social Computing (CSCW), Baltimore, Maryland, February 2014.

[Nature 2013] Nature, Special Issue on the ENCODE project, 6 September 2012.

[NEON 2014] NEON. National Ecological Observatory Network. From http://www.neoninc.org. Last accessed October 7, 2014.

[Nielsen 2011] “Reinventing Discovery.” Nielsen M. Princeton University Press, 2011.

[Nonaka and Takeuchi 1995] Nonaka and H. Takeuchi, The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation. New York: Oxford University Press, 1995.

[Oleksik et al 2014] "Study of an Electronic Lab Notebook Design and Practices that Emerged in a Collaborative Scientific Environment." Gerard Oleksik, Natasa Milic-Frayling and Rachel Jones. Computer Supported Cooperative Work and Social Computing (CSCW), Baltimore, Maryland, February 2014.

[Park et al 2014] "Toward crowdsourcing micro-level behavioral annotations: the challenges of interface, training, and generalization." S. Park, P. Shoemark, and L.-P. Morency. IUI 2014.

[Pietras and Coury 1994] "The Development of Cognitive Models of Planning for Use in the Design of Project Management Systems." C. M. Pietras and B. G. Coury. International Journal of Human-Computer Studies, Vol 40, 1994.

[Polanyi 1983] M. Polanyi, Tacit Dimension. Gloucester, Mass.: Peter Smith Publisher Inc, 1983.

[Raban et al 2010] Daphne R. Raban, Mihai Moldovan, Quentin Jones. "An empirical study of critical mass and online community survival." Proceedings of the ACM conference on Computer supported cooperative work, 2010.

[Ribes and Finholt 2009] "The long now of infrastructure: Articulating tensions in development." Ribes, D. and T. A. Finholt. Journal for the Association of Information Systems (JAIS): Special issue on eInfrastructures 10(5): 375-398, 2009.

[Smith and Goodman 1984] "Understanding Written Instructions: The Role of an Explanatory Schema." E. Smith and L. Goodman. Cognition and Instruction, 1(4), 1984.

[Spinellis and Louridas 2008] Diomidis Spinellis and Panagiotis Louridas. "The Collaborative Organization of Knowledge." CACM, August 2008.

[Steichen et al 2013] "User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities." B. Steichen, G. Carenini, and C. Conati. IUI 2013.

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

[Takeuchi and Nonaka 2004] H. Takeuchi and I. Nonaka, Hitotsubashi on Knowledge Management. Singapore: John Wiley & Sons (Asia), 2004.

[Tausczik et al 2014] "Collaborative Problem Solving: A Study of MathOverflow." Yla R. Tausczik, Aniket Kittur and Robert E. Kraut. Computer Supported Cooperative Work and Social Computing (CSCW), Baltimore, Maryland, February 2014.

[Van Merrienboer et al 2003] "Taking the load of a learners' mind: Instructional design for complex learning." Van Merriënboer, J.J.G., Kirschner, P.A., & Kester, L. Educational Psychologist, 38(1), 2003. I.