A Task-Centered Framework for Computationally-Grounded Science Collaborations

Yolanda Gil, *Member, IEEE*, Felix Michel, *Member, IEEE*, Varun Ratnakar, Matheus Hauder, *Member, IEEE,* Christopher Duffy, Hilary Dugan, and Paul Hanson

***Abstract*—Collaboration is ubiquitous in today's science, yet there is limited support for coordinating scientific work. The general-purpose tools that are typically used (e.g., email, shared document editing, social coding sites), have still not replaced in-person meetings, phone calls, and extensive emails needed to coordinate and track collaborative activities. Scientists with diverse knowledge and skills around the globe could collaborate by opening scientific processes that expose all tasks and activities publicly to achieve a shared scientific question. This paper describes the Organic Data Science framework to support scientific collaborations that revolve around complex science questions that require significant coordination, entice contributors to remain engaged for extended periods of time, and enable continuous growth to accommodate new contributors as the work evolves over time. We discuss how the design of this framework incorporates principles followed by successful on-line communities. We present initial results to date of several communities that are collaborating using this framework.**

***Index Terms*— Computer interfaces, collaborative work, social computing.**

I. Introduction

SCIENTIFIC collaborations, sometimes referred to as "collaboratories" and "virtual organizations", range from those that work closely together and others that are more loosely coordinated [1, 2]. 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
- *Engaging unanticipated participants*, so that the collaboration can 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. 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 initial results of 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 design of the Organic Data Science framework is based on social principles derived from studies of successful on-line communities and collaborative projects where members work closely together towards a common goal. There have been many studies of on-line communities [3], notably on Wikipedia. Our work builds on the social design principles uncovered by this research, as well as other projects that have successful close collaborations on-line. The design of the user interface incorporates those social design principles and to support features that target specific aspects of collaborative work [4, 5]. The framework is an extension of a semantic wiki platform, where the semantic properties of tasks and other entities (people, datasets, software) are used to organize the content and the activities in the collaboration [6]. The Organic Data Science framework is being used by several communities. This paper presents an initial analysis of their characteristics, such as the growth of the communities, the interactions occurring in them, and the kinds of activities that the users are engaging in. The framework is still under development, and it evolves to accommodate user feedback and to incorporate new collaboration features.

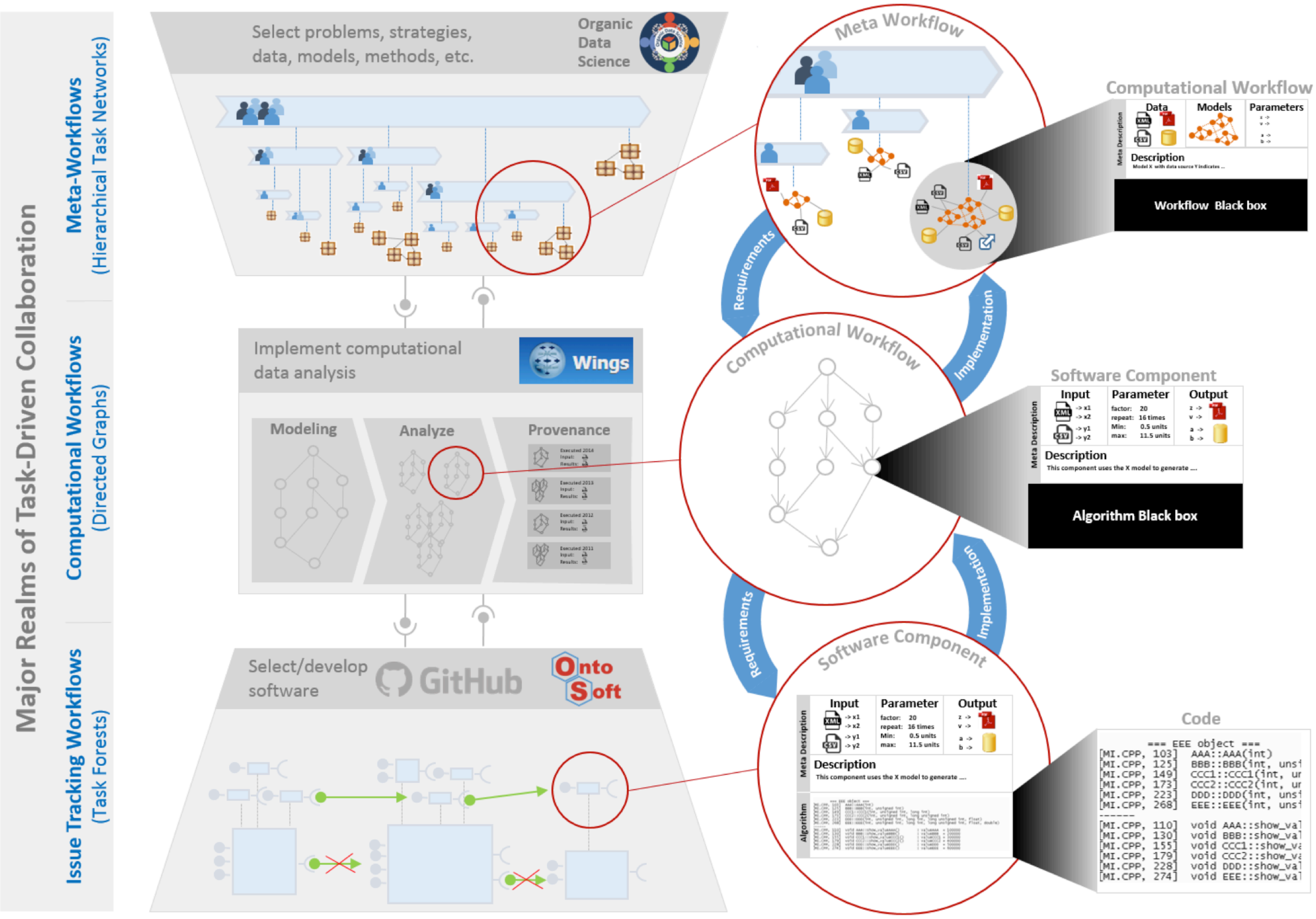
Y. Gil and V. Ratnakar are with the Information Sciences Institute at the University of Southern California, USA (e-mail: gil@isi.edu, varur@isi.edu).

F. Michel was with the Information Sciences Institute at the University of Southern California, USA. He is now with the Department of Software Engineering for Business Information Systems, Technical University of Munich, Germany (e-mail: michelf@in.tum.de).

M. Hauder is with the Department of Software Engineering for Business Information Systems, Technical University of Munich, Germany (e-mail: hauder@in.tum.de).

C. Duffy is with the Department of Civil and Environmental Engineering at Penn State University (e-mail: cxd11@psu.edu).

H. Dugan and P. Hanson is with the Center for Limnology at the University of Wisconsin Madison, USA (e-mail: pchanson@wisc.edu).



```
          === EEE object ===
[MI.CPP, 103]  AAA::AAA(int)
[MI.CPP, 125]  BBB::BBB(int, unsi
[MI.CPP, 149]  CCC1::CCC1(int, ur
[MI.CPP, 173]  CCC2::CCC2(int, ur
[MI.CPP, 223]  DDD::DDD(int, unsi
[MI.CPP, 268]  EEE::EEE(int, unsi
-------
[MI.CPP, 110]  void AAA::show_val
[MI.CPP, 130]  void BBB::show_val
[MI.CPP, 155]  void CCC1::show_va
[MI.CPP, 179]  void CCC2::show_va
[MI.CPP, 228]  void DDD::show_val
[MI.CPP, 274]  void EEE::show_val
```

Fig. 1. Major realms of computationally-grounded collaboration.

The main contributions of this paper are the social and community aspects of our Organic Data Science framework. First, it describes in detail the social design principles that guide its design. Second, it provides an initial report on the collaborative activities of a few communities that are currently using it.

The paper starts with a discussion of the kinds of collaboration that we target, introducing the notion of a *meta-workflow* in scientific collaborations and how collaborative meta-workflows are supported in an Organic Data Science framework. It then describes the social principles that are the basis for the design of the framework, and how they are incorporated into our current implementation. Finally, it presents an initial analysis of its use in several science communities.

II. Layers of Collaboration

Collaboration in science is ubiquitous and occurs at many levels. We are specifically interested in supporting collaboration that is grounded on computational aspects of data science. That is, collaboration activities whose goal is to develop computational methods for data analysis in order to answer science questions that require assembling data, software, and expertise that come from a diverse group of scientists. We refer to this kind of collaboration as *computationally-grounded collaboration*. A key feature of this kind of collaboration is that it is driven by goal-oriented activity, that is, the collaboration has a purpose to accomplish a joint computational goal, and the collaborative activities have an end result that is a solution to accomplish that goal.

We have observed computationally-grounded collaboration in three different realms, illustrated in Figure 1:

1. *Workflow creation activities*: In our prior work on computational workflows [34, 35], we observed that workflows often amalgamate the expertise of several scientists, and are often collaboratively developed. For example, some scientists may have skills on how to interpret and integrate particular kinds of data, others on statistical techniques to do quality control of the data (e.g., data that is coming from noisy sensors), others may understand how to apply a particular physics model, and others may contribute statistical techniques to detect particular types of patterns in the data. These activities result in a computational workflow that implements a data analysis method. These activities are generally supported by workflow systems [7, 8, 9], although the collaborative aspects of workflow design are often not explicitly supported [10]. In our work, we use the WINGS workflow system[1] and have extended it to support the

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

exchange of partially-developed workflows across users, but more work needs to be done to integrate issue tracking and versioning. Workflow creation activities are illustrated in the middle portion of the figure. Computationally speaking, workflows can be represented as directed graphs, and the dependencies expressed in the workflow tend to mirror the dependencies among the scientists involved in the collaboration.

2. *Coding activities*: A separate realm for collaboration is in the development of software. In this realm, scientists work on finding relevant software that may have been previously developed, extending a particular implementation, resolving bugs, and developing new code to address a specific need. These activities are illustrated in the bottom of the figure. Code sharing sites (e.g., GitHub[2]) are well suited to support these kinds of activities. In separate research, we are designing complementary scientist-centered software registries to support some of these activities, in particular for geosciences software as part of the OntoSoft project[3]. The resulting codes are used to implement some part of an overall workflow. These collaborative activities may be top-down when driven by a target workflow, and bottom-up when the existence of software determines the feasibility and design of a workflow. In this realm, the activities are generally driven by separate issues that are eventually resolved, so computationally speaking the collaborative activities tend to resemble a task forest of what we would call issue-tracking workflows.
3. *Meta-workflow design activities*: A third realm for collaboration focuses on the activities that lead to the requirements for the workflows themselves. Examples of such activities in this context include agreeing to a joint question, developing strategies to address that question, figuring out whether there is data and software available to address that question, investigating how to get that data and software, and positing alternative versions of the question to workaround inaccessible data or software. These activities can be seen as meta-workflow design activities, because they lead to sketching out a high-level workflow (or workflows) to address that question (or questions) which is then implemented as a computational workflow. Alternatively these activities may be guided bottom-up when the availability of a workflow may prompt scientists to pose a question that can be answered by that workflow. These meta-workflow design activities are illustrated at the top of the figure. Computationally speaking, these activities tend to be organized as a hierarchical task network [11], where each task is decomposed into subtasks assigned to subsets of those involved in the collaboration [12, 13, 14].

This paper focuses on supporting scientific collaborations that involve meta-workflow design activities. These activities are generally not supported appropriately by collaboration tools, and tend to be done through general-purpose platforms such as email and shared document editing. They also tend to rely heavily in in-person communication. Therefore, scientific collaborations tend to be very time-consuming and have a very high communication cost. Reducing the effort needed for this level of collaboration could allow scientists to do research faster. We also believe that scientists are often deterred from undertaking complex research questions that would result in overhead that is unmanageable with the general-purpose platforms that they currently use.

We believe that the separations across these realms are due in part to the diverse nature of the shared goals and activities, but more importantly they are often separate activities because the tools that support them are not well integrated. A more fluid flow of information across these levels would also improve the efficiency of scientific collaborations.

III. Organic Data Science

Computationally-grounded collaboration occurs at several levels, from high-level meta-workflow design to determine what scientific problem to solve and how, to workflow creation to select the data and analytic software to be used, to coding activities to implement the software needed. The focus of this work is the former, that is, the collaboration that occurs when scientists are working together to agree on a problem to solve and a strategy to solve it. Eventually, a workflow is chosen with appropriate data and software, and run to obtain results that address the original problem. A challenging aspect of these collaborations is that they are often supported by general tools that are not well integrated. For example, a common situation would be that some discussions take place over emails, others through shared on-line documents, and others occur face to face. Another challenge is that different members of the collaboration participate in different activities, making it hard for everyone to have the information they need. For example, the more senior people with the broader vision for a project may participate in face to face meetings only, while post-doctoral researchers may be more involved in the emails and other routine discussions with limited visibility on the general strategy, and the students that do the detailed work do not have much understanding of the project outside of their particular scope of work. This situation makes collaborations very inefficient: there can be misunderstandings (e.g., some solution may be implemented that does not take into account some longer later goal), communication overhead (e.g. having to transmit to different students at different stages in different organizations), and limited experiences for the younger members of the collaboration (in terms of the credit they get, the training they receive, etc). Finally, these collaborations are often across institutions and span long periods of time, making it hard for everyone to track what is happening, who is doing what and when, and most importantly to remember exactly what was done once the work is ready for publication. This makes it especially challenging for newcomers to come up to speed and understand how to contribute. The documentation for these collaborations is scattered in everyone's notes and emails (some people may have moved on and may not be available to provide details), which severely limits the ability of the group and certainly of others to reproduce and build on the work done.

[2] http://www.github.com
[3] http://www.ontosoft.org

We are developing a novel approach called *organic data science* to support these meta-workflow design collaborations. Organic data science captures all the activities, their participants, and associated documents in an open framework that is centered on tasks. Everyone involved in the collaboration has visibility on all other activities and can contribute to them as needed. Newcomers can view what tasks are being pursued by the collaboration and quickly come up to speed and contribute.

The Organic Data Science framework supports these collaborations. The framework is implemented as an extension of a semantic wiki [15] to organize all the information relevant to the collaboration [6]. The framework is heavily influenced by principles extracted from social science studies of diverse on-line collaborations, whether scientific in nature or not, particularly those driven by joint tasks and goals. These principles drive the design of the user interface [4]. They also support the virtual community by fostering self-organization, sustainability, and open dissemination [5]. The next section describes these principles in detail.

IV. Social Design Principles for Organic Data Science

There are numerous studies about successful on-line communities [3]. Many studies are focused on Wikipedia and other wiki-style frameworks, with topics as varied as the design of the editorial process [16], community composition and activities [17], incentives to contributors [18, 19], critical mass of contributors [20], coordination across contributions [21], group composition [22], conflict [23], trust [24], and user interaction design [25]. These studies suggest a number of principles for the design of our on-line collaboration framework.

Figure 2 summarizes the social principles that we are using in our approach. We follow the organization used in [3], 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). Additional social principles are outlined in Figure 3 and represent the best practices and lessons learned from two projects that are applicable to our work. The rest of this section describes briefly all these principles to motivate the design of the system and the communities around it.

Starting Communities

Starting a community is challenging, and many on-line communities never take off. First, the community must co-habitate with the ecosystem of already existing sites. There are lots of web sites relevant to any given area, so it is important to identify the particular niche that the ODS community will cover and describe its scope up front relating it to other sites. The scope should be described in terms of topics to be covered, target members, activities, and purpose. The ODS site cannot be isolated, instead its content should be integrated with those other related sites when possible. Second, members and content should be organized into subspaces in order to facilitate the formation of communities of interest, form their identity based on the content they contribute, and facilitate their interactions. Third, the planned timespan of activities should be clearly marked and active tasks should be brought to the forefront, so that activity in the community can be easily conveyed to a visitor. Activities that are planned for the future should be annotated with the expected timeframe for their activation and target end dates, so members understand the overall plans for the community activities. Finally, creating mechanisms to match people to ongoing activities, perhaps by suggesting to them subspaces where colleagues with similar interests are participating. These principles have been found to facilitate the creation of a critical mass for jumpstarting a core community.

Encouraging Contributions through Motivation

Once there is a critical mass of core members, motivating contributors to add content becomes critical. First, the ODS site should point out to contributors what is needed, for example by highlighting what content is needed or by asking specific people for concrete content based on what they have contributed before. Second, carving out smaller tasks makes it easier for people to volunteer. Large or challenging tasks should be decomposed into smaller ones so they are more achievable piecemeal by different contributors. Specifying the expected end date for a task also helps convey the scope of the commitment to the contributors considering taking it on. Second, positive feedback and encouragement go a long way. Requests for contributions coming from leaders of the project are most effective. Frequent feedback about the value of the contributions is also helpful, particularly if it is positive and not just guidance or critiques. Concrete rewards for accomplishments, i.e., not just for signing up but for finishing a task), are also very effective even if very small or intangible. Third, peer pressure is very effective. Publicizing what others have accomplished and that they complied with their commitments sets a certain tone in the collaboration that stimulates contributions. Finally, people are more likely to contribute if they understand that their personal expertise is needed for the task and they have a commitment to the success of the group.

Encouraging Commitment

The sustainability of the community is important, so strategies for encouraging long-term commitment are crucial to the success of an on-line community. Helping people connect as a subgroup helps connect individuals to the community by identifying with that subgroup. Subgroups should have their own identity, e.g., a name or tagline, and a clear relationship to the larger group. The goals of a subgroup should be clear with respect to the goals of the overall community. Their purpose should be also explicit, so the subgroup is not just an abstract entity but has reasons to interact and work together towards a common goal. Finally, interdependent tasks increase commitment and at the same time reduce conflict among contributors.

Attracting and Engaging Newcomers

The sustainability of a community is also ensured through its growth. Therefore, attracting and engaging newcomers is crucial to the success of an on-line community. First, a most effective way to engage new members is to have current members approach their colleagues. Second, there should be a point person(s) appointed to have the first interactions with

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

Fig. 2. Selected social principles from [3] for building successful online communities that can be applied to the Organic Data Science framework. 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

Fig. 3. Selected best practices from the Polymath project [26] and lessons learned from ENCODE [27] that guided the design of our Organic Data Science framework.

new members, addressing their questions and helping until they see some examples of how things work. Introducing them to other members also helps them to be engaged. Third, some initial barriers should be put in place, just to ensure that if a newcomer overcomes those barriers they are inclined to contribute in principle and the investments made in them will not be wasted. Fourth, announcements about new members should be disseminated to the community, with pictures and personal background information. Announcements coming from community leaders help new members understand where direction for the collaboration is coming from and how. Fifth, providing incentives to early members is important since many of the principles just mentioned are harder to accomplish for a community of very small size. Sixth, designing common learning experiences for newcomers helps them feel that they have earned their right to be part of the community and they have been given training so they can feel empowered to contribute from the beginning. Offering sandboxes where they can do initial practices is also very helpful. It also helps to have a clear articulation of community rules, and the stages for newcomers to go through in order to become full-fledged members. Throughout these stages, members should be given more privileges and control over the system.

Best Practices from Polymath

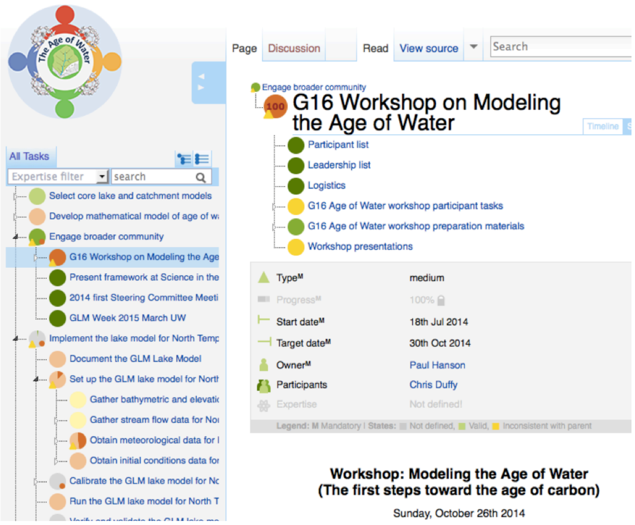
We find inspiration in the Polymath project, set up to collaboratively develop proofs for mathematical theorems [26, 28], where professional mathematicians collaborate with volunteers that range from high-school teachers to engineers to solve mathematics conjectures. The collaboration is centered around 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 [29] with wiki pages that are used for write-ups of basic definitions, proof steps, and overall final publication [30]. 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 [31]. 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.

Lessons Learned from ENCODE

Another project that has exposed best practices of a large collaboration is ENCODE [32, 27]. 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.

V. Collaborating with the Organic Data Science Framework

Our Organic Data Science framework is designed to incorporate the social principles described in the prior section. Figure 4 shows a snapshot of the user interface, illustrating task decomposition (left and top right), task metadata (center right) such as participants and start/end times, and task documentation (bottom right) which includes semantic properties of tasks (not shown in the figure). We describe elsewhere how each of the social design principles influenced the design of specific features of the user interface [4, 5], and how the framework extends and uses a semantic wiki platform [6], in particular Semantic MediaWiki [15], to allow users to create structured representations of tasks and other entities (datasets, people, software) relevant to the collaboration.



- Select core lake and catchment models
- Develop mathematical model of age of w
- Engage broader community
  - G16 Workshop on Modeling the Age
  - Present framework at Science in the
  - 2014 first Steering Committee Meeti
  - GLM Week 2015 March UW
- Implement the lake model for North Temp
  - Document the GLM Lake Model
  - Set up the GLM lake model for Nortl
    - Gather bathymetric and elevatic
    - Gather stream flow data for Noı
    - Obtain meteorological data for l
    - Obtain initial conditions data for
  - Calibrate the GLM lake model for Nc
  - Run the GLM lake model for North T

Engage broader community

G16 Workshop on Modeling the Age of Water

- Participant list
- Leadership list
- Logistics
- G16 Age of Water workshop participant tasks
- G16 Age of Water workshop preparation materials
- Workshop presentations

Type^{M}	medium
Progress^{M}	100%
Start date^{M}	18th Jul 2014
Target date^{M}	30th Oct 2014
Owner^{M}	Paul Hanson
Participants	Chris Duffy
Expertise	Not defined!

Legend: M Mandatory | States: ▨ Not defined, 🟩 Valid, 🟨 Inconsistent with parent

Workshop: Modeling the Age of Water
(The first steps toward the age of carbon)

Sunday, October 26th 2014

Fig. 4. Sample task page from the AoW community.

The software is open source[4] and can be forked on GitHub to create a new community. New created communities should be registered to create a central list of existing communities.

Communities

Several communities are currently using the Organic Data Science framework. The major use of our framework is by a community of hydrologists and limnologists that are studying the age of water (AoW) in an ecosystem. This involves determining the concentrations of water isotopes at different locations as water flows over time. They are the main driver for the development of the Organic Data Science framework, and the AoW site initially included activities that span both topics. As the community evolved, it was eventually split into two separate sites (AoW and ODSF).

Another community is the ENIGMA consortium for neuroimaging genetics[5]. This consortium includes more than 70 institutions that collaborate to do joint neuroscience studies. The institutions keep their data locally, but they all agree to the method and software to be used to analyze their data. They organize themselves into working groups, each group studies a particular disease (e.g., autism) and cohort (e.g., children.) The leads of the consortium are interested in using the Organic Data Science framework to track what institutions participate in what study, the characteristics of their datasets, and the point person in that institution for each particular study. A requirement of this group is that some information needs to remain private to outsiders, and other information can only be shared between each institution and

[4] https://github.com/IKCAP/organicdatascience
[5] http://enigma.ini.usc.edu/publications/the-enigma-consortium-in-review/

the lead organization. As a result, they have set up two separate sites: ENIGMA-LEADS and ENIGMA-ALL. Both sites are referred to here as ENIGMA.

Another community is a group of geoscientists working together to publish a special issue of a journal composed of geoscience papers of the future (GPF). All the articles will follow a similar format in that they publish explicitly all datasets, software, and workflows used to generate the results in the paper. The site is being used to coordinate the activities involved in tracking the status of each paper, and to compare the approaches in different papers.

We have also set up a site for training new users of the Organic Data Science framework regardless of their home community (ODST). All new users are given a pre-defined set of tasks each involving learning about some aspect of the framework. This gives them the ability to use this new system in a practice setting, following one of the social principles described earlier for newcomers. As they practice, they can create their own tasks and add themselves as participants of other tasks.

In the rest of the section we present a brief and preliminary analysis of the initial data that we have available from these communities that are still in initial formation stages. The data is sensitive in nature and is not publicly released. The software to gather, analyze, and visualize data for an Organic Data Science site is part of the framework software release[4].

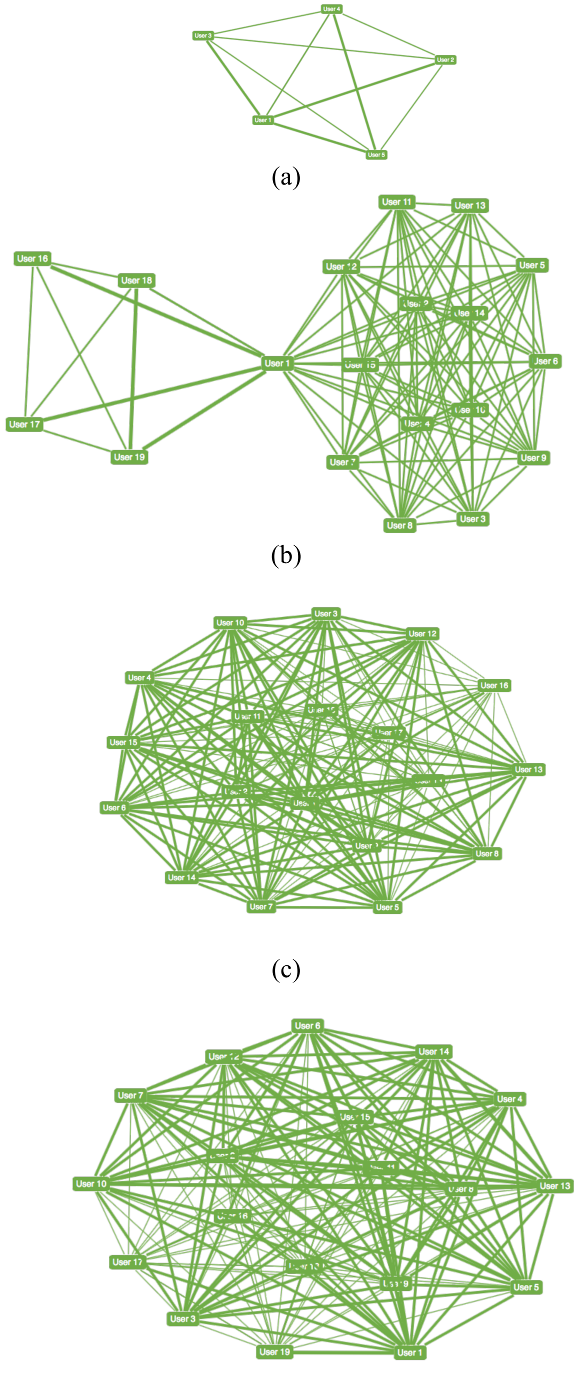
Preliminary Analysis

Each Organic Data Science community has a dashboard that is publicly accessible and shows aggregate data about the collaboration. It includes collaboration graphs generated from the task metadata properties that link tasks and users. We call these graphs *social task networks* [33], where each user is a node in the graph and the links indicating whether two users have a task in common (i.e., being either owner or participant). The thickness of a link indicates how many tasks the two users have in common.

Figure 5 illustrates the evolution of one of the communities (GPF) by showing the social task network at four different points in time. The GPF community was seeded with five organizers of the special issue (3a). The organizers shared different sets of tasks involved in planning the special issue. One of the organizers served as the host for the authors (3b). The authors shared more and more tasks as the collaboration progressed (3c). Eventually, the members of the community shared different amounts of tasks (3d), so the thickness of the lines is more pronounced in the final graph. We describe in [Gil et al 2015b] the evolution of the AoW and ODSF communities, starting from a single site where two distinct subgraphs can be seen in the social task network and later two distinct (but overlapping) communities working in two separate sites.

Figure 6 shows some of the data about the tasks. The tasks hierarchies tend to be shallow (3 to 5 nodes deep) and have in some cases hundreds of nodes. The figure also shows the social task network, which illustrates the number of users and the strength of the connections among them through the number of tasks they share. Note that all the sites have many more users, but they are not shown here because they are not so strongly connected to others through their tasks.

Table 1 gives a summary of the metadata for the tasks in the different communities. We show the summary for all the tasks, and then we show the data for tasks that have incomplete metadata (participants, owner, start/end dates, the type, and expertise required) and then the tasks with complete metadata. The average length of the tasks, the amount of participants per task, and the task type (indicating high- mid- or low-level task) vary across the different collaborations.



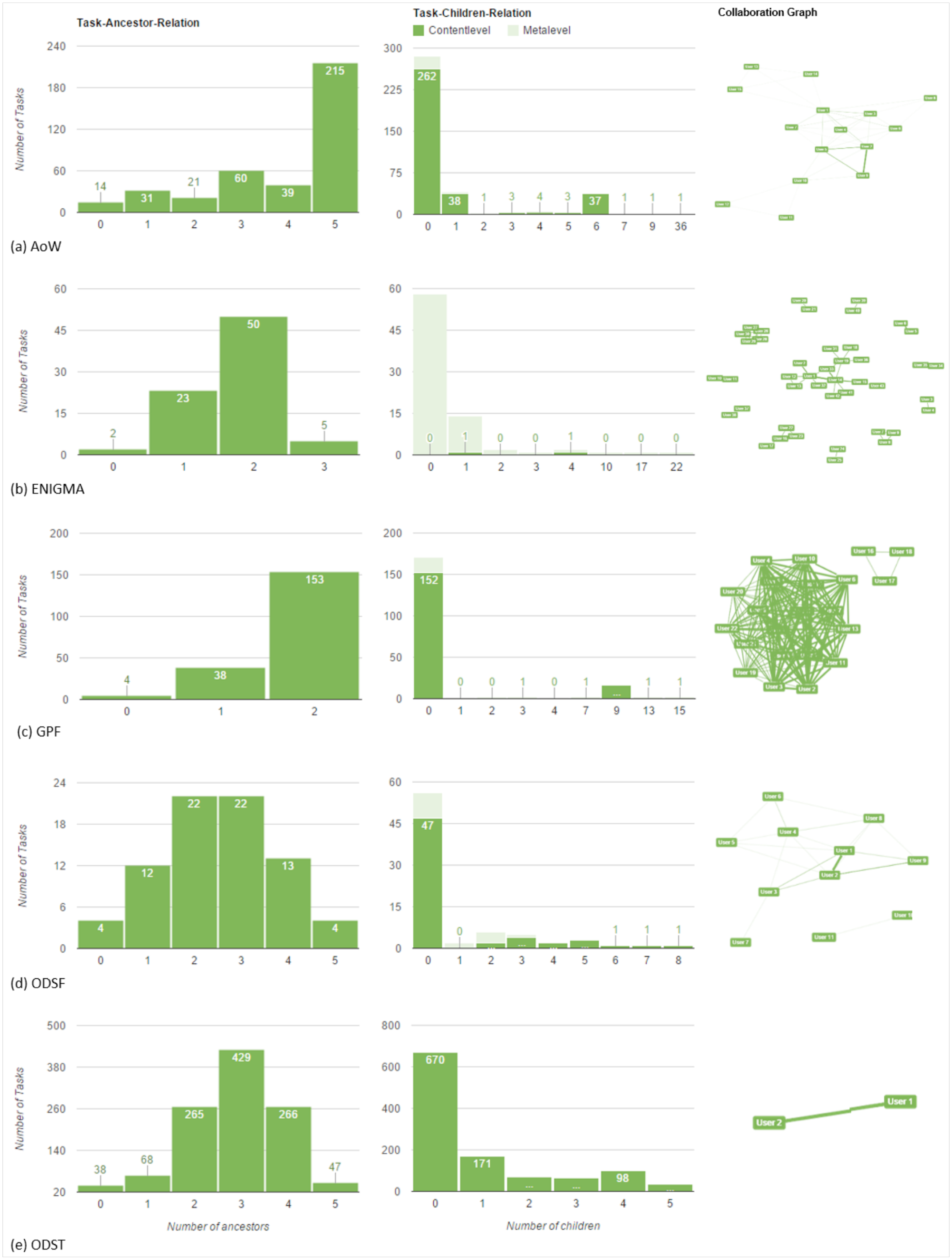
(a)

(b)

(c)

(d)

Fig. 5. Evolution of the collaboration in the GPF community



(a) AoW

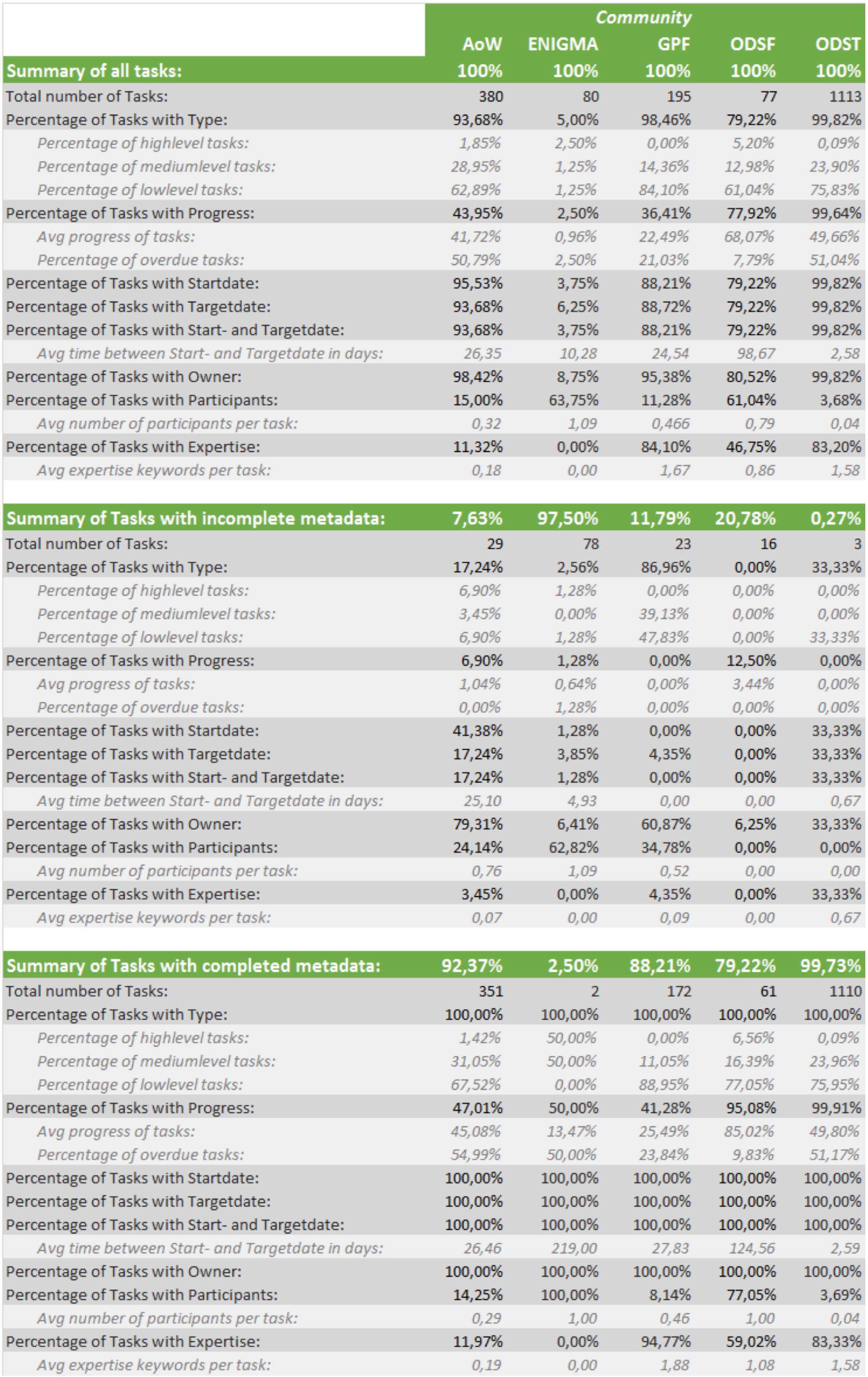
(b) ENIGMA

(c) GPF

(d) ODSF

(e) ODST

Fig. 6. Characteristics of tasks in different Organic Data Science communities.



			Community		
	AoW	ENIGMA	GPF	ODSF	ODST
Summary of all tasks:	**100%**	**100%**	**100%**	**100%**	**100%**
Total number of Tasks:	380	80	195	77	1113
Percentage of Tasks with Type:	93,68%	5,00%	98,46%	79,22%	99,82%
Percentage of highlevel tasks:	*1,85%*	*2,50%*	*0,00%*	*5,20%*	*0,09%*
Percentage of mediumlevel tasks:	*28,95%*	*1,25%*	*14,36%*	*12,98%*	*23,90%*
Percentage of lowlevel tasks:	*62,89%*	*1,25%*	*84,10%*	*61,04%*	*75,83%*
Percentage of Tasks with Progress:	43,95%	2,50%	36,41%	77,92%	99,64%
Avg progress of tasks:	*41,72%*	*0,96%*	*22,49%*	*68,07%*	*49,66%*
Percentage of overdue tasks:	*50,79%*	*2,50%*	*21,03%*	*7,79%*	*51,04%*
Percentage of Tasks with Startdate:	95,53%	3,75%	88,21%	79,22%	99,82%
Percentage of Tasks with Targetdate:	93,68%	6,25%	88,72%	79,22%	99,82%
Percentage of Tasks with Start- and Targetdate:	93,68%	3,75%	88,21%	79,22%	99,82%
Avg time between Start- and Targetdate in days:	*26,35*	*10,28*	*24,54*	*98,67*	*2,58*
Percentage of Tasks with Owner:	98,42%	8,75%	95,38%	80,52%	99,82%
Percentage of Tasks with Participants:	15,00%	63,75%	11,28%	61,04%	3,68%
Avg number of participants per task:	*0,32*	*1,09*	*0,466*	*0,79*	*0,04*
Percentage of Tasks with Expertise:	11,32%	0,00%	84,10%	46,75%	83,20%
Avg expertise keywords per task:	*0,18*	*0,00*	*1,67*	*0,86*	*1,58*
Summary of Tasks with incomplete metadata:	**7,63%**	**97,50%**	**11,79%**	**20,78%**	**0,27%**
Total number of Tasks:	29	78	23	16	3
Percentage of Tasks with Type:	17,24%	2,56%	86,96%	0,00%	33,33%
Percentage of highlevel tasks:	*6,90%*	*1,28%*	*0,00%*	*0,00%*	*0,00%*
Percentage of mediumlevel tasks:	*3,45%*	*0,00%*	*39,13%*	*0,00%*	*0,00%*
Percentage of lowlevel tasks:	*6,90%*	*1,28%*	*47,83%*	*0,00%*	*33,33%*
Percentage of Tasks with Progress:	6,90%	1,28%	0,00%	12,50%	0,00%
Avg progress of tasks:	*1,04%*	*0,64%*	*0,00%*	*3,44%*	*0,00%*
Percentage of overdue tasks:	*0,00%*	*1,28%*	*0,00%*	*0,00%*	*0,00%*
Percentage of Tasks with Startdate:	41,38%	1,28%	0,00%	0,00%	33,33%
Percentage of Tasks with Targetdate:	17,24%	3,85%	4,35%	0,00%	33,33%
Percentage of Tasks with Start- and Targetdate:	17,24%	1,28%	0,00%	0,00%	33,33%
Avg time between Start- and Targetdate in days:	*25,10*	*4,93*	*0,00*	*0,00*	*0,67*
Percentage of Tasks with Owner:	79,31%	6,41%	60,87%	6,25%	33,33%
Percentage of Tasks with Participants:	24,14%	62,82%	34,78%	0,00%	0,00%
Avg number of participants per task:	*0,76*	*1,09*	*0,52*	*0,00*	*0,00*
Percentage of Tasks with Expertise:	3,45%	0,00%	4,35%	0,00%	33,33%
Avg expertise keywords per task:	*0,07*	*0,00*	*0,09*	*0,00*	*0,67*
Summary of Tasks with completed metadata:	**92,37%**	**2,50%**	**88,21%**	**79,22%**	**99,73%**
Total number of Tasks:	351	2	172	61	1110
Percentage of Tasks with Type:	100,00%	100,00%	100,00%	100,00%	100,00%
Percentage of highlevel tasks:	*1,42%*	*50,00%*	*0,00%*	*6,56%*	*0,09%*
Percentage of mediumlevel tasks:	*31,05%*	*50,00%*	*11,05%*	*16,39%*	*23,96%*
Percentage of lowlevel tasks:	*67,52%*	*0,00%*	*88,95%*	*77,05%*	*75,95%*
Percentage of Tasks with Progress:	47,01%	50,00%	41,28%	95,08%	99,91%
Avg progress of tasks:	*45,08%*	*13,47%*	*25,49%*	*85,02%*	*49,80%*
Percentage of overdue tasks:	*54,99%*	*50,00%*	*23,84%*	*9,83%*	*51,17%*
Percentage of Tasks with Startdate:	100,00%	100,00%	100,00%	100,00%	100,00%
Percentage of Tasks with Targetdate:	100,00%	100,00%	100,00%	100,00%	100,00%
Percentage of Tasks with Start- and Targetdate:	100,00%	100,00%	100,00%	100,00%	100,00%
Avg time between Start- and Targetdate in days:	*26,46*	*219,00*	*27,83*	*124,56*	*2,59*
Percentage of Tasks with Owner:	100,00%	100,00%	100,00%	100,00%	100,00%
Percentage of Tasks with Participants:	14,25%	100,00%	8,14%	77,05%	3,69%
Avg number of participants per task:	*0,29*	*1,00*	*0,46*	*1,00*	*0,04*
Percentage of Tasks with Expertise:	11,97%	0,00%	94,77%	59,02%	83,33%
Avg expertise keywords per task:	*0,19*	*0,00*	*1,88*	*1,08*	*1,58*

Table. 1. Analysis of task metadata for different Organic Data Science communities.

The average time between start and target date of tasks is approximately 32 days. Boundary values occur in the trainings wiki where the average task time is only 2.6 days and in the framework wiki where the average task time is approximately 99 days. We believe that the high difference between the communities is related to the evolution of every community. The task structure of a community evolves over time, at the beginning many siblings' tasks are created that are grouped later into more abstract tasks with a deeper nested structure, the average task time increases.

VI. Conclusions

This paper presented the social aspects of the Organic Data Science framework to support computationally-grounded scientific collaboration focused on meta-workflow design that leads to computational workflows. We discussed the social design principles coming from studies of on-line collaboration that we found relevant to this kind of scientific collaboration. The paper also presented preliminary data on the different communities that are currently using the framework. These data show that the collaborations are active and the communities are growing over time.

In future work, we plan to do a formal evaluation to assess how the framework supports scientific collaboration and whether it increases productivity and community growth. We continue to improve and extend the framework based on new requirements and feedback from the different communities.

Acknowledgements

We gratefully acknowledge support from the US National Science Foundation with awards IIS-1344272 and ICER-1440323. We would like to thank other members of the Organic Data Science project, in particular Jordan Read. We would also like to thank the early adopters of the Organic Data Science framework for their feedback and comments on this work.

References

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

[2] N. Bos, A. Zimmerman, J. S. Olson, J. Yew, J. Yerkie, E. Dahl, and G. M. Olson. "From Shared Databases to Communities of Practice: A Taxonomy of Collaboratories." Journal of Computer-Mediated Communication 12(2): 652-672 (2007).

[3] R. E. Kraut, and P. Resnick "Building Successful Online Communities: Evidence-Based Social Design.". MIT Press, 2011.

[4] F. Michel, Y. Gil, V. Ratnakar, and M. Hauder. "A Task-Centered Interface to On-Line Collaboration in Science." Proceedings of the ACM Conference on Intelligent User Interfaces (IUI), 2015.

[5] F. Michel, Y. Gil, and M. Hauder. "A Virtual Crowdsourcing Community for Open Collaboration in Science Processes." Submitted to the Americas Conference on Information Systems (AMCIS), 2015.

[6] Y. Gil, F. Michel, V. Ratnakar, J. Read, M. Hauder, C. Duffy, P. Hanson, and H. Dugan. "Supporting Open Collaboration in Science through Explicit and Linked Semantic Description of Processes." Proceedings of the European Semantic Web Conference (ESWC), 2015.

[7] E. Deelman, G. Singh, M. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, B. Berriman, J. Good, A. Laity, J. Jacob, and D. Katz. "Pegasus: A framework for mapping complex scientific workflows onto distributed systems." Scientific Programming, 13(3), 2005.

[8] B. Ludaescher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao. "Scientific workflow management and the Kepler system." Concurrency and Computation: Practice and Experience. Volume 18. 2006.

[9] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and Huy T. Vo. "VisTrails: Visualization meets Data Management." Proceedings of ACM SIGMOD 2006.

[10] J. Freire, and C. Silva. "Towards Enabling Social Analysis of Scientific Data." Proceedings of CHI Social Data Analysis Workshop, 2008.

[11] M. Ghallab, Da. Nau, and P. Traverso. Automated Planning: Theory & Practice. Morgan Kaufmann, 2004.

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

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

[14] J.J.G. Van Merriënboer. "Training Complex Cognitive Skills: A Four-Component Instructional Design Model for Technical Training.“ Educational Technology Pubns 1997.

[15] M. Krötzsch, and D. Vrandecic. Semantic MediaWiki. Foundations for the Web of Information and Services 2011: 311-326.

[16] D. Spinellis, and P. Louridas. "The Collaborative Organization of Knowledge." Communications of the ACM, August 2008.

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

[18] A. Mao, E. Kamar, Y. Chen, E. Horvitz, M. E. Schwamb, C. J. Lintott, and A. M. Smith. "Volunteering Versus Work for Pay: Incentives and Tradeoffs in Crowdsourcing." HCOMP 2013.

[19] J. Leskovec, D. Huttenlocher, and J. Kleinberg. „Governance in Social Media: A case study of the Wikipedia promotion process." Proceedings of the International Conference on Weblogs and Social Media, 2010.

[20] D. R. Raban, M. Moldovan, and Q. Jones. "An empirical study of critical mass and online community survival." Proceedings of the ACM conference on Computer supported cooperative work, 2010.

[21] A. Kittur, B. Lee, and R. 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.

[22] S. K. Lam, J. Karim, and J. 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.

[23] A. Kittur, and R. E. Kraut. "Beyond Wikipedia: coordination and conflict in online production groups." Proceedings of the 2010 ACM conference on Computer supported cooperative work, 2010..

[24] D. L. McGuinness, H. Zeng, P. Pinheiro da Silva, L. Ding, D. Narayanan, and M. Bhaowal. "Investigations into Trust for Collaborative Information Repositories: A Wikipedia Case Study." Workshop on Models of Trust for the Web, 2006.

[25] R. Hoffmann, S. Amershi, K. Patel, F. Wu, J. Fogarty, and D. S. Weld. "Amplifying community content creation with mixed initiative information extraction." Proceedings of the 27th international conference on human factors in computing systems, 2009.

[26] M. Nielsen. "Reinventing Discovery." Princeton University Press, 2011.

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

[28] T. Gowers. "Is Massively Collaborative Mathematics Possible?" Retrieved 23 July 2015 from http://gowers.wordpress.com/2009/01/27/is-massively-collaborative-mathematics-possible

[29] M. Nielsen. "The Polymath Wiki." Retrieved 23 July 2015 from http://michaelnielsen.org/polymath1.

[30] T. Gowers. "The Polymath Project.“ Retrieved 23 July 2015 from http://polymathprojects.org

[31] T. Gowers. "General Polymath Rules." Retrieved 23 July 2015 from http://polymathprojects.org/general-polymath-rules

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

[33] Y. Gil, V. Ratnakar, T. Chklovski, P. Groth, and D. Vrandecic. "Capturing Common Knowledge about Tasks: Intelligent Assistance for To Do Lists." ACM Transactions on Interactive Intelligent Systems, 2(3). 2012.

[34] Y. Gil, V. Ratnakar, J. Kim, P. González-Calero, P. Groth, J. Moody, and E. Deelman. "WINGS: intelligent workflow-based design of computational experiments." IEEE Intelligent Systems 26 (1). 2011.

[35] Y. Gil, P.A. Gonzalez-Calero, J. Kim, J. Moody, and V. Ratnakar. "A Semantic Framework for Automatic Generation of Computational Workflows Using Distributed Data and Component Catalogs." Journal of Experimental and Theoretical Artificial Intelligence, 23 (4), 2011.