Bridging the data divide

A workshop on advancing an industry/academic partnership model for Open Collaboration Research

March 15, 2015
Vancouver, BC

Organized by
Jonathan T. Morgan
Aaron Halfaker
Dario Taraborelli
Sean Goggins
Tim Hwang

Hosted by
The 18th Annual ACM conference
on Computer-Supported Cooperative Work
and Social Computing
(CSCW 2015)

Manuscript prepared by Jonathan T. Morgan, Aaron Halfaker, Sean Goggins, and Paul J. Weiss

Version 1.1

Details of the workshop, agenda and all documentation can be found here:
https://meta.wikimedia.org/wiki/Research:CSCW'15_workshop
Contact: jmorgan@wikimedia.org

Cite this work as:
Morgan, Jonathan T.; Halfaker, Aaron; Taraborelli, Dario; Goggins, Sean; Hwang, Tim (2015): Open Collaboration Systems Research Workshop 2015 Report.
http://dx.doi.org/10.6084/m9.figshare.1577641

OCS Workshop 2015 Report

Overview

Abstract

A group of researchers, designers, and community managers who work within the domain of Open Collaboration Systems (OCSs) met for a day-long workshop at the 2015 Computer-Supported Cooperative Work conference in order to outline a set of practical, methodological, and theoretical challenges that impact our ability, as a community, to understand and support open collaborations effectively.

During the workshop, the group defined a set of four Research Directions *that our community can take to help focus our work going forward. Each of these research direction addresses a different set of considerations for collaborating across institutional and disciplinary boundaries to further a set of common goals:*

- **Theoretical questions** *that pertain to OCSs generally and have relevance for both basic and applied OCS research*
- **Analytical dimensions** *across which OCSs and OCS research activities can be compared to generate new knowledge*
- **Design requirements** *for shared infrastructure to support high quality, open, reproducible, and collaborative OCS research*
- **Collaboration strategies** *for maintaining productive and mutually-beneficial collaborations between academic and industry stakeholders in OCS research*

The group also outlined a set of concrete, actionable next steps that different OCS stakeholders can perform to address the challenges and opportunities identified across the four research dimensions.

Participants

- Krystle Cheung - WikiHow
- Giovanni Luca Ciampaglia - Indiana University at Bloomington
- Kevin Crowston - Syracuse University
- Jeremy Foote - Northwestern University
- Michael Gilbert - University of Washington
- Sean Goggins - University of Missouri
- Aaron Halfaker - Wikimedia
- Benjamin Mako Hill - University of Washington
- Tim Hwang - Imgur
- Stuart Lynn - Zooniverse
- Alan McConchie - Stamen Design, University of British Columbia

- David McDonald - University of Washington
- Jonathan T. Morgan - Wikimedia
- Gabriel Mugar - Syracuse University
- Jeff Nickerson - Stevens Institute of Technology
- Carsten Østerlund - Syracuse University
- Felipe Ortega - Universidad Rey Juan Carlos
- Sneha Narayan - Northwestern University
- Toby Negrin - Wikimedia
- Aaron Shaw - Northwestern University
- Dario Taraborelli - Wikimedia
- Katherine Thornton - University of Washington
- Paul J. Weiss - University of Washington
- Andrea Wiggins - University of Maryland

Introduction

Open online communities like Wikipedia, OpenStreetMap, WikiHow, Zooniverse, and Imgur support self-organized, cooperative work by distributed networks of volunteers. Both CSCW researchers and industry practitioners who support these open collaboration systems (OCSs) communities seek to build understanding about the nature of open collaboration. However, there are few opportunities for academic and industry researchers to communicate and collaborate.

On March 14, 2015 in Vancouver, British Columbia we brought together stakeholders across the OCS research ecosystem to discuss the state of open collaboration research and practice, and to develop recommendations for advancing the field of OCS research and improving outcomes for OCS creators, contributors, users, and community organizers. We developed a program around two broad questions:

- What are the most pressing questions pertaining to our scientific understanding of OCSs?
- How can we improve collaboration and data-sharing between academia and industry?

The workshop was held during the annual Computer-Supported Cooperative Work and Social Computing (CSCW) conference of the Association for Computing Machinery (ACM). CSCW was chosen as a venue because a large volume of high quality research on OCSs is published in its annual proceedings.

Purpose

This document is intended to provide scholars, designers, managers, and communities with a set of relevant, current considerations for OCS research and practice agreed upon by leading researchers in the field. This document can serve as a reference point for future research, to

direct new work and help justify it. OCS stakeholders may use it to identify knowledge gaps, research opportunities, and design directions, and to argue both the academic merit and practical utility of specific new initiatives that were called for in this workshop.

Process

The program consisted of a series of ~45 minute breakout sessions interspersed with lightning talks and roundtable discussions. During each breakout, a team of 4-5 participants discussed issues related to the workshop's themes (Access and Theory), taking notes on a dedicated etherpad. Throughout the day, select participants from OCS organizations and academic institutions gave 6-8 minute talks that focused on their research (in the case of academics) or the history and dynamics of the OCSs they support (in the case of industry professionals).

During 'Access' breakouts, teams focused on challenges that various stakeholders (e.g.. professors, research scientists, community contributors) face related to data sharing, as well as strategies for facilitating more effective collaboration, proposals for best practices, and a frank discussion of incentive structures within their organizations.

During 'Theory' breakouts, teams assessed the state of OCSs research, examining issues such as the potential benefits and limitations of current theoretical approaches, and the generalizability and practical applicability of findings from the current body of research, published in ACM conference proceedings and elsewhere.
After each round of breakout sessions, one member of each team summarized the team's key findings and questions for the rest of the workshop participants, who had the opportunity to question and comment during an informal roundtable discussion. During the afternoon breakouts, teams focused on synthesizing and prioritizing their notes from the morning session into 2-3 main points.

After the workshop a self-selected group of organizers and participants performed an additional synthesized the five groups' notes, and grouped them together under four thematic headings (Practical theory, Comparative analysis, Shared research infrastructure, and Collaborations). The writers then proposed a set of concrete steps for all members of the OCS research community, based on the issues and proposals raised during the workshop. We believe that the considerations, priorities, and recommendations presented here represent a basic consensus among a diverse set of stakeholders that can help build and sustain a community of practice around OCSs that spans individuals, institutions, disciplines, communities, and technologies.

Research directions

In this section, we provide a summary of recommendations for the future of the field that emerged from the workshop. These recommendations are provided to highlight perceived gaps in our current understanding of OCSs or in our approach to OCS research and design, to direct attention to emerging research opportunities, and to help develop a consensus around "best

practices" for collaborative work in our community. The recommendations are grouped according to four general themes: *Practical theory, Comparative analysis, Shared research infrastructure*, and *Collaborations.*

Practical theory

There is a substantial gap between theory of OCSs and practice. While "A/B tests" are common in industry, they tend to be used "atheoretically", or only to address immediate questions about a specific software change. Academics work to push theory forward, though not necessarily to get their design recommendations implemented. As a result, design and evaluation of living OCS systems is seldom informed by theory.

Further, the academic distance between disciplines that tend to study different types of OCS makes generalization difficult. (Wiki, Open science, Open source, etc.) For example, papers that discuss the nature of Wikipedia seldom reference and integrate related work studying open source software or other crowdsourced projects. This has resulted in silos of proto-theory that are ripe for tearing down.

Theory-based research on OCSs should allow for comparative analysis and integration of findings across work contexts/communities/platforms. We identified three key areas of concern that would benefit immediately from more multidisciplinary, theory-driven research.

How can we socialize new volunteers in mature, complex OCSs?

With increasing norm complexity, it seems that this open socialization pattern stops working so well.[1][2] Attempts at formalizing mentor/mentee relationships have been unsuccessful -- largely due to the small scale at which the project operated and due to the difficulty of entrance for mentees.[3] Experiments in introducing new means of socialization support have shown promise qualitatively, but improvements on newcomer retention have not been experimentally demonstrated.[4][5][6]

How can we make software changes without disrupting work practices?

Recently managers of OCSs have struggled to deploy improvements to mature OCS without causing substantial disruptions. For example, nearly every major software change that the Wikimedia Foundation has deployed to large, mature Wikipedia communities has been surrounded in conflict that pits the "staff" vs. the "volunteers" (e.g. new contribution mechanisms[1], WYSIWYG editor[2] and new media viewer[3]). Recent literature suggests that this is not a phenomenon that is limited to wiki medium (e.g. [7]). The CSCW literature has a celebrated history of calling attention to invisible infrastructures of socio-technical work practice[8], but there's a lack of practical advice for managers of mature OCSs. Is there a way that software changes can be designed, developed and deployed in a way that minimizes this disruption?

What is a "normal" life-cycle for an OCS?

English Wikipedia is arguably the largest, mature, online open collaboration projects that has been studied in the CSCW literature. Recent work has explored declining trends in the

community's volunteer population[9] that seem to have been caused by increasing rule complexity and the need for efficient quality control processes[2]. While it's clear that this trend has lead to decreased productivity, it's not clear whether this is "normal" or even "healthy" for an OCS of this scale. If such declines in participation are undesirable, what strategies will allow the communities to recover? Addressing these questions is of critical importance to both the maintainers and community leaders of large scale OCS.

Comparative Analysis

Performing comparative analysis between OCSs is difficult for many reasons, including incompatible data formats, inconsistent access to data, and incongruent metrics for activity, quality, contribution, etc. Most research studies focus on a single site, at a single point in time. This has lead to a lack of generalization: while some work has found evidence for 'strong regularities' in participation, but it is not clear, for instance, how to compare the work of power users in creating and curating WikiHow manuals and StackExchange Question threads.

Practical considerations govern many academic researchers' decisions of which OCS to study. A large portion of academic work to date uses data from English Wikipedia because the data is open and the WMF hosts individuals who can act as liaisons. While Wikipedia is a rich and multifaceted data source, this focus on Wikipedia likely colors our understanding of OCSs generally. It is not clear, for example, whether Wikipedia's exponential growth trajectory, followed by an extended plateau in participation, is characteristic of OCSs in general. This may be preventing advancement of the field and leading to ill-advised design decisions or misleading conclusions when dealing with other (smaller, non-wiki, differently-focused) OCS projects.

Further, many studies draw from a single theoretical framework, using one type of methodology or analysing an OCS at only one level. The tendency is understandable given the distribution of OCS researchers across myriad disciplines and departments, and the depth of expertise required to perform rigorous, theory-driven research. However, it has resulted in a lack of integrated theories about the nature of OCS and the siloing of the literature into different journals and conference proceedings.

Finally, "least publishable unit" approach to research, common in fields like HCI, encourages the proliferation of one off studies that limit their scope to a single moment in the lifespan of OCSs, rather than longitudinal studies focused on temporal shifts, trends, and cycles of activity.
To address these challenges, we call for an ecological approach to OCS research that examines systems across multiple dimensions.

Comparison across time

More longitudinal research is required in order to increase our understanding of the trajectories, shifts, and cycles of activity over the lifespan of an OCS.

Comparison across OCSs

More comparative research is required in order to increase our understanding of the influence of factors such as the size of a system, the type of work, and community dynamics on the character of an OCS. One underdeveloped area of research that we consider to be particularly important is analysis of what constitutes success and failure in different OCSs.

Comparison across theories and epistemologies

In order to develop a coherent body of theory around open collaboration, we must understand the advantages and drawbacks of applying theoretical frameworks that were developed elsewhere. This involves critically assessing the suitability of theories and metaphors borrowed from other domains of science, such as biological ecology, organization science, and economics for OCS research. It will also be necessary to assessing the commensurability of findings from theory-driven studies with different epistemological bases (for example positivism vs. interpretivism).

Comparison between levels of analysis

Open collaborations are complex socio-technical systems. Understanding these systems involves examining individual components (people, software features), aggregates (communities, platforms), as well as the interactions between the system and its context--the cultures, institutions, and infrastructures with which the system and its human participants are embedded. Comparing the explanatory power of theories that work at different levels of analysis (for example theories of individual motivation vs. social interaction vs. system dynamics), and evaluating the compatibility of research methods that operate at different analytical levels is of critical importance.

We recognize that current incentive structures for OCS researchers and practitioners often conspire against open data access, interdisciplinary research, and generalizability and reproducibility of research findings.

However, incremental progress can be made. Other organizations can learn from the Wikimedia Foundation's model of releasing regular, open licensed database dumps and providing access to a full history via a live API. This strategy has encouraged many researchers to study Wikimedia Projects, and resulted in a wealth of actionable research findings, novel methods for data analysis, and innovative design prototypes.

Conferences and journals that seek to publish high quality OCS research can also learn from venues that have adopted more open approaches to review and access. Public Library of Science (PLOS) has been a pioneer in open access, with a review process that focuses on methodological rigor rather than disciplinary conformity. By doing away with explicit page limits and instituting a staged review process, the ACM Computer Supported Cooperative Work conference (CSCW) has demonstrated that it is possible to maintain high quality standards and

a relatively high (~30%) acceptance rate while relaxing arbitrary constraints around the size of publishable “unit” of research.

Shared research infrastructure

Shared infrastructure for Open Collaboration Systems research is necessary now for a number of reasons, all of which center on a) leverage advantages of shared infrastructure, and b) intellectual cohesiveness advantages of shared infrastructure.

Both leverage and cohesiveness across disciplines are enabled by a shared infrastructure because OCSs are fundamentally distinct networking and collaboration phenomena that facilitate the collective construction of tangible or intangible products using flexible, distributed, and non-hierarchical forms of organization. The emergence of widely available, highly flexible, interactive information infrastructure technologies significantly altered the universe of feasible organization structures and strategies. OCSs represent a new class of organizing solutions, in which individuals self-organize in order to collaboratively produce any number of artifacts and experiences. OCSs differ from other popular online structures, such as crowdsourcing platforms or online social networks in significant ways. In crowdsourcing, the firm or client proposing the project typically controls the decision-making process. In online social networks, organized collective production is not usually a goal for participants. The evolution and potential of OCS structures and processes creates a need for the type of interdisciplinary community building proposed. OCSs now play a critical role across the many domains of society.

Developing a more coherent understanding of OCSs through the development across datasets would further multiple intellectual disciplines, and yield economic, health and educational benefits for societies that are becoming more technologically mediated. A community of scholars dedicated to a more coherent unpacking of the success patterns, failure patterns and factors affecting growth and performance in OCSs will make concrete contributions to businesses, governments, and citizens. The current diffuseness of OCS research makes discipline specific findings difficult for society to utilize because most organizations lack the incentives, time, and capacity to parse and prioritize each scientific discipline’s unique perspective on OCSs. Presently, OCS researchers often lack the incentives, knowledge, and capacity to create truly sharable data resources. This communication and coherence gap around knowledge construction in OCSs limits the impact of OCS research and development.

This proposed project is a first, but important step toward building a community that overcomes existing obstacles.

Direct collaboration

- We also identified strategies that can be used while infrastructure is in development:
- Omnibus survey & cross-community surveys
    - Surveys perform many roles for community and researchers -- work together on common questions

- collaboration on projects/grant proposals
    - HOWEVER… Direct collaboration between industry and academia is difficult due to mismatched timescales of organizational needs, grants and open-ended work.
    - If academics and industry practitioners are aligned on problems/needs, then they should take advantage of the opportunity to work together.
    - In these cases, it's easier to justify engineering and research support for academics and write that support into grant proposals.

Next steps

In this section, we describe the immediate next steps participants of the workshop and other OCS research should take. We've grouped the items by who should be responsible for implementing them: industry practitioners, academics or both in collaboration.

For industry researchers

Formal recognition of research support

Industry organizations can reap substantial benefits from external researchers. Spending even a small amount of resources on research support can lead to substantial value. For example mailing lists like wiki-research-l[10] and DERP proposals[11] have developed communities that provide a routing service to help researchers find datasets and related work around a platform. These lists can be developed and supported by a single person or small group of people part-time. Advocate for formal recognition of research liaison and advisory roles & responsibilities within your organization.

Release your data

Make data available in ways that researchers need. This means releasing historical datasets and publishing APIs that make historical analysis possible.

Document for an academic audience

Researchers will be more likely to study your community if your resources are well documented. Document your data sources, your processes/policies, and your "big questions" for researchers to reference in papers and grant proposals.

For academics researchers

Write for an industry audience

Make your research findings available to relevant industry practitioners and articulate their relevance to immediate problems that organizations face.

Formal recognition for collaboration

Academic fields stand to benefit from improved collaboration between industry and academic. While this means the work ought to qualify as service work, it doesn't fit into the current models. New models of service work will need to be developed around supporting collaborative research/data infrastructure maintenance.

Leverage your relationship with an OCS

Approach OCS orgs with funding collaboration opportunities (e.g., NSF grants). While industry practitioners may not need the grant's funding they may be willing to endorse and otherwise support projects that align with their goals.

For all stakeholders

Cross-community studies and grant proposals

Gather researchers who have built expertise in different communities/datasets to study common phenomena (e.g. http://arxiv.org/abs/1411.2878). This is a temporary solution to the lack of shared technical infrastructure and common documentation.

Collaborate on shared infrastructure

Both academics and industry researchers benefit from better platforms for sharing research proposals, datasets and results. Define roles and share responsibilities for the maintenance of this infrastructure. Use this platform to develop standard dataset/API formats and prototypical datasets.

Conclusion

While this workshop allowed us to come together, strengthen our community and start a call for new work in access and theoretical development, there's still much work that will need to be coordinated. Future workshops should either focus directly addressing one of the next steps called out above (e.g. collaborating on shared infrastructure) or produce an expanded set of concrete next steps.

References

1. Heather Ford and R. Stuart Geiger. 2012. "Writing up rather than writing down": becoming Wikipedia literate. In Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration (WikiSym '12). ACM, New York, NY, USA, Article 16 , 4 pages. DOI=10.1145/2462932.2462954 http://doi.acm.org/10.1145/2462932.2462954

2. Halfaker, A., Geiger, R. S., Morgan, J. T., & Riedl, J. (2012). The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. American Behavioral Scientist, 0002764212469365.

3. Musicant, D. R., Ren, Y., Johnson, J. A., & Riedl, J. (2011, October). Mentoring in Wikipedia: a clash of cultures. In Proceedings of the 7th International Symposium on Wikis and Open Collaboration (pp. 173-182). ACM.

4. Morgan, J. T., Bouterse, S., Walls, H., & Stierch, S. (2013, February). Tea and sympathy: crafting positive new user experiences on wikipedia. In Proceedings of the 2013 conference on Computer supported cooperative work (pp. 839-848). ACM.

5. Aaron Halfaker, R. Stuart Geiger, and Loren G. Terveen. 2014. Snuggle: designing for efficient socialization and ideological critique. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14). ACM, New York, NY, USA, 311-320. DOI=10.1145/2556288.2557313 http://doi.acm.org/10.1145/2556288.2557313

6. Von Krogh, G., Spaeth, S., & Lakhani, K. R. (2003). Community, joining, and specialization in open source software innovation: a case study. Research Policy, 32(7), 1217-1241.

7. Gazan, R. (2011, May). Redesign as an act of violence: disrupted interaction patterns and the fragmenting of a social Q&A community. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2847-2856). ACM.

8. Grudin, J. (1988, January). Why CSCW applications fail: problems in the design and evaluation of organizational interfaces. In Proceedings of the 1988 ACM conference on Computer-supported cooperative work (pp. 85-93). ACM.

9. Suh, B., Convertino, G., Chi, E. H., & Pirolli, P. (2009, October). The singularity is not near: slowing growth of Wikipedia. WikiSym (p. 8). ACM.

10. https://lists.wikimedia.org/mailman/listinfo/wiki-research-l

11. http://derp.institute/#collaborate