Advancing the OCDX: Building Social Computing Infrastructure

Sean Goggins

University of Missouri
gogginss@missouri.edu

Matt Germonprez

University of Nebraska-
Omaha
mgermonprez@unomaha.edu

A.J. Million

University of Missouri
ajmillion@gmail.com

Aaron Halfaker

Wikimedia Foundation
ahalfaker@wikimedia.org

Libby Hemphill

Illinois Institute of Technology
libbyl.hemphill@itt.edu

Abstract

We propose a workshop that builds on prior CSCW workshops focused on building a shared technical, social, and methodological infrastructure for more transparently sharing: a) data connected to social computing papers, b) methods and software built to analyze that data, and c) a common set of ethical practices built to address issues emerging in social computing research.

Author Keywords

Open collaboration, CSCW tools, Data factories, Metadata, Research methods.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction

Open online communities (OOC) have emerged as significant drivers of innovation, economic activity, and social well-being. OOCs play important roles in a wide variety of areas, including but not limited to software development, general knowledge management, education, health, and scientific discovery. Scholars and practitioners from different disciplines (e.g., computer science, sociology, mathematics, economics,

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). Copyright is held by the author/owner(s).
CSCW'17 Companion, February 25–March 1, 2017, Portland, OR, USA.
ACM ISBN 978-1-4503-4676-4/17/02.
<http://dx.doi.org/10.1145/3022198.3022661>

physics, anthropology, organization science, and communications) engage in OOC research to build disciplinary understanding that in turn helps citizens successfully manage and grow OOCs. For example, management scholars in free and open source software (FOSS) focus on developing theories of collaboration on these projects drawn from rich, qualitative methods [6], while software engineering scholars address developer coordination tools [3] and specific issues of how to make sense of electronic trace data through software repository mining [1, 2]. HCI scholars in FOSS are particularly focused on how tools might be designed to support different modes of collaboration [4]. The research contexts are identical, but differences in data and method prevent the development of coherent understandings across these disciplines. Similar challenges exist in studies of Wikipedia, online health support communities, citizen science projects, open online courses, and a myriad of specific OOCs. Building a community of scholars who address differences in research aims, data and method offers the potential to enable a new, interdisciplinary synthesis of OOC knowledge. In turn, this will increase the coherence of scientific and public communication across disciplines that study OOCs [7].

OCDX Project Target Deliverables

Building *prototypes* of valid, reliable, replicable, large-scale, high-impact studies of OOC's requires a computing and policy community across disciplines to build a sustainable vision for the collection and generation of open collaboration data. *There is a strong need to promote publication of details about the whole process, not just the analysis and results.* Toward this end, the Open Collaboration Data Factory (OCDX)

(<http://www.ocdx.io/>) proposes three deliverables that individually provide a **foundation for future collaborations** among OOC researchers: (1) datasets and analysis workflows that can serve as models [5], vetted by the CSCW community in part, through this workshop, (2) new data sets and analysis workflows of types with proven utility for researchers, to be developed by the OCDX core group, and (3) model policies and procedures for OOC data resource creation, for use by Institutional Review Boards (IRB) and stakeholders addressing policy development around privacy in OOCs. Together, the proposed deliverables will create a model for dissemination of technology and data, including policies and guidelines for dataset preparation, documentation, and certification. The resulting prototype tools, policies, and vision will be connected to a number of published, scholarly papers, white papers, and technical reports.

Accomplishments to Date, Identifying Gaps Leading to Supplement Request

OCDX accomplishments to date have included building international working groups and developing prototype infrastructure. After successfully building an international virtual organization through workshops at CSCW 2014, 2015 & 2016; through a symposium in Copenhagen, and workshops and hackathons at ACM Group 2014 and 2016, the OCDX successfully:

- Developed an OOC infrastructure survey to send to stakeholders at universities, in government, and in industry. This survey was intended to identify stakeholder needs for the OCDX and the recently developed OCDX manifest.

- Planned and executed the OCDX Data Manifest workshop from January 27-29, 2016 in Omaha, Nebraska. This workshop refined prior versions of the OCDX manifest descriptive framework and set the stage for future activities.
- Hosted a workshop at CSCW on February 27, 2016 in San Francisco, California. Wikimedia Foundation supporters indicating continued industry support and were responsible for submitting the proposal. In this workshop, we began to prototype an end-to-end infrastructure called for in an OCDX proposal. Therefore, in this workshop, we experimented with documentation protocols and technologies that were designed to make the process of “breaking into” a new dataset more tractable for researchers studying open online communities. This workshop’s purpose was to bring together researchers to test these systems and discover problems and missed opportunities to support iteration.

Prior successes identified two critical resource need categories, leading us to call for this workshop with the aims of: a) increasing engagement from newly identified international partners and to b) enhance the infrastructure available for a Wizard of Oz Prototype piloted at CSCW 2016.

Building Out OCDX Ingestion Software and Process Pilot

The OCDX has developed the OCDX manifest that forms the core of the originally proposed Wizard of Oz prototype. From June 2016 through May 2017 the

OCDX continued to validate and refine the prototype. Creating a small pilot test environment requires additional resources, however, including cloud infrastructure and technologist and software developer time that was not originally anticipated. Consequently, this workshop will help us to engage the CSCW community to advance our work via:

- Scaling tooling to enable the production and discovery of the existing OCDX manifests.
- Advancing design and development efforts related to the OCDX (this will include a small user-needs and user-experience oriented workshop that will be supported through funds from the original award that have not yet been spent).

Our current efforts and infrastructure development efforts are available on GitHub at <http://www.github.com/OCDX>.

References

1. B. Al-Ani, Erik Trainer, Roger Ripley, Anita Sarma, Andre van der Hoek, and DF Redmiles. 2008. *Continuous Coordination within the Context of Cooperative and Human Aspects of Software Engineering*. ACM, Leipzig, Germany.
2. C. Bird, P.C. Rigby, E.T. Barr, D.J. Hamilton, D.M. German, and P. Devanbu. 2009. *The promises and perils of mining git*. IEEE.
3. Kelly Blincoe, Giuseppe Valetto, and Sean Goggins. 2012. Proximity: a measure to quantify the need for developers’ coordination. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 1351–1360.

4. L. Dabbish, C. Stuart, J. Tsay, and J. Herbsleb. 2012. *Social Coding in Github: Transparency and Collaboration in an Open Software Repository*. ACM, Seattle, Washington.
5. Sean Goggins and Christopher Mascaro. 2013. Context Matters: The Experience of Physical, Informational and Cultural Distance in a Rural IT Firm. *The Information Society* In Press.
6. James Howison and Kevin Crowston. 2014. Collaboration Through Open Superposition: A Theory of the Open Source Way. *Mis Quarterly* 38, 1: 29–50.
7. Nora McDonald and Sean Goggins. 2013. Performance and participation in open source software on github. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, 139–144.