Workshop: Designing and Building a Space for Innovative Scholarly Practices to Enhance Open Online Community Scholarship

http://www.ocdx.io

https://github.com/ocdx

ABSTRACT

Open online communities play important roles in a wide variety of areas including, but not limited to, software development, knowledge management, education, health, and scientific discovery. Finding ways to bring scholars together to discuss diverse interests and abilities promotes dataset curation and fosters coherent disciplinary understandings of scholarship needs. This abstract discusses the "Open Collaboration Data Factory," which is a virtual research institute and sociotechnical space designed to foster interdisciplinary OOC research.

CCS Concepts

Heterogeneous (hybrid) systems; Programming interfaces; Collaborative and social computing

Keywords

Data management; Open online communities; Interdisciplinary scholarship

1. INTRODUCTION

Researching open online communities is central to the HCI literature, and it is not out of the question that, frankly, we have been doing a lot of things wrong. For example, it is unlikely that any two papers from different labs examining Wikipedia, GitHub, eBird, Facebook, Twitter or any other space where open online communities emerge, will have a) clear descriptions of the provenance of their data; b) open access to scripts and anonymized samples of the data; c) complete methods descriptions; or d) consistency between them, even if the research questions are similar. As scientists, we should recognize that this state of affairs represents a hole in our process; but not a spiritual hole that we should be aware of accept, but a scientific hole we should actively seek to close. The project described here, the "Open Collaboration Data Factories" project is one effort to close this hole.

Open online communities (OOCs) are fundamentally distinct phenomena that facilitate the collective construction of 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. OOCs represent a new class of organizing solutions, in which individuals self-organize in order to collaboratively produce any number of artifacts and experiences. OOCs 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 OOC structures and processes creates a need for more coherent approaches to interdisciplinary research, as evidenced by the number of CHI papers on these topics.

CHI itself is interdisciplinary, but we are not as much a "self contained" "interdiscipline" as we are one among a sea of growing interdisciplinary buoys; or perhaps floating planks from the shipwrecks of traditional disciplines. Spanning intellectual disciplines is potentially risky. Developing new interdisciplinary practices and methodological approaches could conflict with a discipline's current discourse and findings. However, the potential benefits of interdisciplinary OOC work is significant for both science, which gains leverage from integrated research models and the corresponding advancement in knowledge; and society, which is growing increasingly reliant upon OOCs. The principle aim of this abstract is to describe the role new mechanisms for scientific collaboration within CHI and between CHI and other "interdisciplines" can play in the development of a new research community, the Open Collaboration Data Factory (OCDF).

Abstract, top down mechanisms for more systematic OOC research are not likely to succeed. Instead, we need to gear our efforts towards problem solving in a flexible and adaptive network of methods, tools and data structures. The authors' experiences with interdisciplinary scholarship suggests that by "doing", research collaborations we are more likely to develop new approaches than by merely talking about them.

Scholars and practitioners from different disciplines (e.g., computer science, sociology, mathematics, economics, physics, anthropology, organization science, communications) engage in research about OOCs from particular disciplinary points of view, which in turn helps citizens successfully manage and grow OOCs in specific ways. 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 [2], and specific issues of how to make sense of electronic trace data through software repository mining [1]. Human computer interaction (HCI) scholars in FOSS are particularly focused on how tools might be designed to support different modes of collaboration [3]. The research contexts are identical (and all these types of studies have been published in CHI proceedings, and elsewhere), but differences in data and method prevent the development of coherent understandings across these disciplines and, arguably, within CHI. Similar challenges exist in studies of Wikipedia, online health support communities, citizen science projects, open online courses and a myriad of specific OOCs.

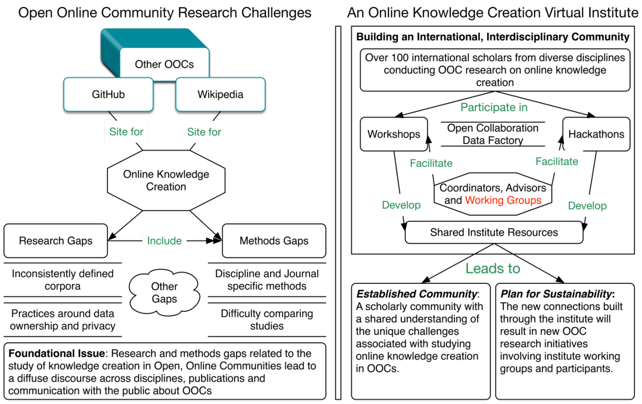
This introduces a second, related challenge, which is the participation of individuals in multiple different communities, and the associated limitations of scholarship focused on only one of those communities. For example, knowledge construction occurs in Wikipedia, GitHub, Facebook and Twitter, but few studies [10] examine behavior across different OOCs.

The work of the OCDF is building a high level "software and data manifest" to enable cross context and interdisciplinary scholarship. The work is low level, abstract, and unsexy; yet essential for the science of examining OOCs. Our initial investigations into different points of view and expectations in the evaluation of high-level descriptive metadata practices surfaced teh unique areas of expertise that OCDF members from different parts of CHI and different scholarly communities can contribute to fuller and better understandings of the structure, contents and future value of OOC datasets.

In some respects we are simply operationalizing a scholarly kumbaya for a challenge many recognize; but its hard to get competitors to agree while also leveraging the "unique advantages" of their lab's research methods in the competitive realm of gaining scholarly publications and being a "first mover". This paradox may not be overcome; but we have reached a point in OOC research where some aspects of data and tooling can be standardized in order to enable a next round of research progress. At some point, are we not competing so heartily amongst ourselves that outside "interdisciplines" might devour our prey?

A lack of mutual recognition of scholarly contributions may stymie innovation and advancements that encourage transparent and progressive communication within and across OOCs. Building a diverse community of scholars who address differences in research aims that are reflected in collection and analysis methodologies will enable a new, interdisciplinary synthesis of OOC knowledge. In turn, this will increase the coherence of scientific and public communication across existing disciplines that study OOCs [5]. This includes, but is not limited to the CHI community.

Figure one illustrates the foundational issue in the study of online knowledge creation in OOCs on the left, and outlines the activities and aims of the proposed a virtual institute on the right. The issues of coherence within different communities studying the same data set makes discussion about a particular OOC difficult across those communities. There is a diffuse discourse across a range of disciplines about knowledge creation in open online communities. In an effort to encourage recognition of these shared interests and practices the authors established an Open Collaboration Data Factory (OCDF), which serves as a virtual institute for discussion and research that promotes transparent and sustainable work on OOCs. For example, projects like the data census seek to identify OOC dataset repositories and metadata practices it has been possible to identify and leverage areas of shared interest and needs (both technological and disciplinary).


Open Online Community Research Challenges
Other OOCs
GitHub
Wikipedia
Site for
Site for
Online Knowledge Creation
Research Gaps
Include
Methods Gaps
Inconsistently defined corpora
Other Gaps
Discipline and Journal specific methods
Practices around data ownership and privacy
Difficulty comparing studies
An Online Knowledge Creation Virtual Institute
Building an International, Interdisciplinary Community
Over 100 international scholars from diverse disciplines conducting OOC research on online knowledge creation
Participate in
Workshops
Open Collaboration Data Factory
Hackathons
Facilitate
Facilitate
Coordinators, Advisors and Working Groups
Develop
Develop
Shared Institute Resources
Leads to

Foundational Issue: Research and methods gaps related to the study of knowledge creation in Open, Online Communities lead to a diffuse discourse across disciplines, publications and communication with the public about OOCs

Established Community: A scholarly community with a shared understanding of the unique challenges associated with studying online knowledge creation in OOCs.

Plan for Sustainability: The new connections built through the institute will result in new OOC research initiatives involving institute working groups and participants.

Figure 1: The complexities of OOC research

Figure one illustrates the complexities of OOC research, while highlighting the benefits of interdisciplinary scholarship that uses OOC data. There is a diffuse discourse across a range of disciplines about knowledge creation in open online communities. The left side of this figure illustrates the foundational issue at the base, and then points to two specific gaps in research and research methods, noting Wikipedia and GitHub as two possible targets for work. The right side of this figure visually outlines the OCDF strategy for building an interdisciplinary, international institute from community of scholars. A central aspect of effort is that bringing people together around one concrete issue for OOC research will build stronger connections. This will foster an institute sustained by small groups of participants that would not otherwise emerge.

The OCDF's interest in fostering collaboration, in this light, provides a framework for giving different types of specialized scholarly knowledge and ability equal weight within a single project. The OCDF builds on existing open data sharing technologies, such as the DataHub (http://datahub.io/) project (sponsored by the Open Knowledge Foundation) and GitHub (http://github.com), to link and document scattered datasets from different OOC. Recognizing the contributions and achievements of researchers and research projects will increase participations and help clarify the value of interdisciplinary collaborations in future projects. It is also facilitating refinement and widespread distribution of available computational resources, code and tools, practical procedures, surveys, reports, open-access publications, and annotated case studies. To overcome obstacles to collaboration, the OCDF is organized by three working groups: Infrastructure (Practice and Practices), Privacy & Ethics and Metadata and Metadata & Ontologies. Primary investigators (PIs) make up a leadership team that coordinates overlapping projects amongst working groups. One project was an open data census.

2. SITUATING OOC RESEARCH AND CURRENT OBSTICLES

The current diffuseness of OOC research within scholarly communities, interdisciplinary or otherwise, makes the practical application of our findings at best difficult. This is because most people, groups and organizations lack the incentives, time, and capacity to parse and prioritize each scientific discipline's unique perspective on OOCs. Presently, OOC researchers often lack the incentives, knowledge, and capacity to create truly sharable data resources. This communication and coherence gap around knowledge construction in OOCs limits the impact of OOC research and development.

Our initial work includes the important, but unglamorous task of organizing discussions among scholars as CSCW, the Consortium for the Science of Sociotechnical Systems, ACM Group and at a standalone symposium at Copenhagen Business School in 2015. These are important steps toward building a community that overcomes existing obstacles.

Developing a more coherent understanding of OOC's through interdisciplinary analyses 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 OOCs will make concrete contributions to *businesses, governments, and citizens.*

The OCDF has begun building a common set of easily accessible and replicable research datasets (http://census.datafactories.org) and processes (http://www.datafactories.org) that support research addressing the design and theoretical challenges raised by diverse multi-community OOC systems. The OCDF is also developing ethical and privacy guidelines for processing OOC data as an integral part of its mission and deliverables. All products and outcomes will be available under open licenses to further promote the reproducibility of studies and methods. Specifically, the focused community developed through this project will design and prototype workflow tools to encourage creation of high-quality datasets; practice guidelines to facilitate consistent application of common research methods; and documented research and resource exemplars to support coherent mapping of theory, data-based measures, and available analytic tools in OOC research. The project will also identify, characterize and annotate available datasets and software tools and identify strategies for increasing compatibility between them in order to promote OOC research built on data from different contexts and platforms.

Each aspect of the OCDF supports making known datasets available; consistently producing OOC related datasets of different types, and developing processes and policies that enable the OCDF to provide data resources that support high-impact studies of OOCs. Identifying and describing existing datasets and tools for OOC dataset analyses establishes a platform for outreach and innovation, which will make best practice guidelines, schema and infrastructures developed by the OCDF more effective. While these interactions will stem from shared interests and modes of communication, it is hoped that sustained partnerships will foster growing understandings of the array of skills and practices that contribute to scholarship about OCCs. That is to say, instead of seeking to normalize scholarly and technical practices, the OCDF hopes to foster increased understanding and appreciation for differing approaches to studying OOCs. The OCDF enables ongoing OOC research projects, both in the US and internationally, to share expertise and resources from different scientific areas to resolve the current fragmentation of OOC research. Furthermore, this OCDF will introduce both early and experienced researchers interested in OOCs to emerging theories, methods, data sources, practical procedures, and tool prototypes for the study of OOCs, facilitating the refinement of research practice in this field and promoting coordination of research efforts across individual groups and institutions.

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, are developing a foundation for future collaborations: (1) existing datasets and analysis workflows that can serve as models, vetted by a community of researchers through workshops and initial use. (2) New data sets and analysis workflows of types with proven utility for researchers, to be developed by the OCDF core group, and (3) model policies and procedures for OOC data resource creation, for use by Institutional Review Boards (IRB) and stakeholder's addressing policy development around privacy in OOCs. The remainder of this paper will discuss how discussions of metadata and indexing needs have helped contextualize the work and interests of OCDF working group members.

In the fall of 2014, the OCDF began to conduct a pilot data census using Semantic MediaWiki. Four instructors at University College Dublin, the University of Maryland, McKendree University, and the Illinois Institute of Technology created assignments requiring undergraduate and graduate students to find and describe OOC datasets in a shared online repository. The result was a collection of datasets meant to aid researchers; however, as graduate research assistants reviewed records created by students it became clear that a number of inconsistencies amongst students' descriptive practices existed. Additionally, other OOC researchers involved with the OCDF had yet to articulate best practices and shared needs for data within the census. The initial data census was an opportunity to identify repositories for OOC datasets and, simultaneously, to evaluate existing metadata practices in the repositories hosting them. Based on initial efforts, graduate research assistants began the process of creating a metadata schema and workflow for the OCDF.

There were two goals for building an OCDF metadata schema and workflow. First was establishing where to collect information about datasets. Second was outlining best practices for metadata creation, revision and maintenance. Next, adopting stakeholder feedback, graduate researchers created a revised metadata schema and preliminary guidelines for adding datasets to the repository. Implementing the schema and workflow has required ongoing communication amongst working group members and potential stakeholders. Treating stakeholders as a user group responsible for the use and creation of OOC data supports the identification of needs associated with dataset metadata. These OCDF stakeholder expectations were to: (1) improve the metadata, (2) clarify census input processes, (3) determine data quality (i.e., what to include), and (4) ascertain the technical infrastructure needed to share data.

3. A MANIFEST FOR OOC DATA

The OCDF's interest in inclusion and collaboration supports the integration of multiple technical and theoretical models for dataset evaluation and documentation. Specifically, interests in balancing the development of infrastructures and technologies with the evaluation and discussion of ethical and concerns framed the need for a metadata schema capable of capturing and reflecting the value and complexity of OOC datasets. The OCDF workflow uses the DCC Curation Lifecycle Model (figure 2) to guide a series of iterative tasks that support the identification of actors, actions and technologies that contribute to the collection, creation and maintenance of records. Based on these iterative tasks it has been possible to foster discussion and collaboration about characteristics of OOC datasets, while simultaneously evaluating methods and technologies for replicating scholarship that uses them.

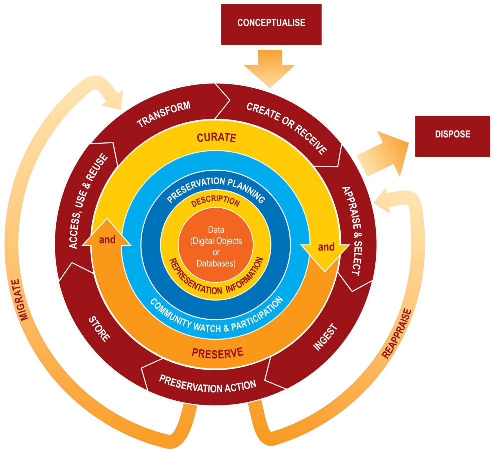


Figure 2: DCC Curation Lifecycle model [7]

The lifecycle model contains six total rings, while each requires a series of action and contributions. Establishing iterative tasks maximizes participant input on the quality and accuracy of specific metadata fields and/or entire metadata records. Using a data curation lifecycle that promotes flexible and ongoing data management has made possible to integrate multiple points of view into the standards for metadata creation and the interface users interact with while creating records for datasets. To augment the coherence of the curation practices expressed in the lifecycle, an additional metadata workflow model was adopted (figure 3).

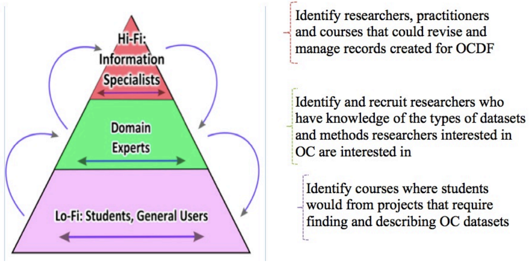

Hi-Fi: Information Specialists
Domain Experts
Lo-Fi: Students, General Users
Identify researchers, practitioners and courses that could revise and manage records created for OCDF
Identify and recruit researchers who have knowledge of the types of datasets and methods researchers interested in OC are interested in
Identify courses where students would from projects that require finding and describing OC datasets

Figure 4: Metadata evaluation and creation workflow [8].

Identifying areas of overlap in metadata creation and/or revision practices created additional space for conversations amongst researchers, which has provided opportunities for different areas of expertise and interest to take priority, but not at the cost of overarching interdisciplinary needs. In general, there are two different goals of establishing a workflow: first, establishing where to collect information about datasets from, and second, outlining best practice guidelines for metadata creation, revision and maintenance.

Taken together the DCC Lifecycle Model and metadata workflow established standards for metadata entry and revision that synthesize research interests and needs while simultaneously recognizing the need for flexibility and diversity in some fields based on disciplinary and technological practices. Developing metadata standards that reflect the interests of researchers from a variety of academic disciplines will enhance the variety and quality of contributions to the OCDF. Keeping these interests and goals in mind led to the development of a manifest, which consists of two documents: a metadata schema and documentation outlining how to use the schema.

The manifest consists of two documents: a schema, and documentation outlining how to use the schema. The schema contains four general descriptive categories (agent, description of dataset, description of data source, metadata creation), and within these four general descriptive areas a series of refinements that facilitate more specific descriptions of documentation, processing and accessibility of a dataset. Documentation outlining how to use the metadata schema provides structure and guidance on the function of each field. Offering guidelines on implementation supports consistent description of OOC datasets, which furthers the OCDF's goal to support interdisciplinary scholarship.

Establishing the structure and purpose of the manifest will support further discuss

4. WORKSHOP TECHNOLOGY

This workshop will be built on an infrastructure and toolset that enables the sharing of electronic trace data from a wide range of systems, including open online community systems, in such a way that the content, structure and associated analysis tooling for each dataset is explicitly noted in an instance of the OCDX manifest. *The proposed manifest will advance the present one by describing the entire research ecosystem around an online behavioral dataset.* Advancing this technical goal makes the analysis of similar online environments and the identification of similar analytical strategies practical and possible for the first time. There are three primary use cases that the proposed infrastructure and toolset will support, with each use case being essential to closing the research transparency gaps that motivate our workshop: (1) ingestion, (2) cross-repository contrasts, and (3) research collaborations (Provenance and Collaboration View). Ingestion happens when a new dataset is integrated into the repository. The ingestion component will read the data file and generate an OCDX manifest instance for that dataset that is as complete as possible. Next, the dataset submitter will fill in the remaining data for the OCDX manifest. Finally, an OCDX manifest file will be created, attached to the dataset, and each will be integrated into the OCDX manifest repository and dataset repository respectively.

Cross-repository contrasts, like comparing Genbank results across labs, remains the holy grail of online community research. As in the case of Genbank, it is the OCDX manifest implementation that makes cross dataset comparisons possible. The complicated, time consuming (and rarely performed) work of identifying similar datasets and drawing comparisons across them will become simple and transparent. The repository engine will become a boundary negotiating artifact that enables technologists, computer scientists, social scientists, and researchers to reference a common process and set of artifacts used to describe, ingest, analyze, and compare online data sets.

To expand on the cross repository contrast idea, imagine that a social scientist has a question about the relationship between the length of tenure of a user in an online community and the content production of that user within that community. A tool that could quickly select all datasets that have those data fields, and potentially even run some simple analysis on the results would facilitate a much larger exploration of research questions and enable the replication of research results, hopefully, resulting in a much more robust set of findings. To enable this process, workshop participants will help to build additional tools on top of and in concert with the OCDX Manifest Engine that includes support for the generation, management, and consumption of OCDX metadata standard derived manifests.

The research collaboration use case is met through a special integration of Jupyter Hub, Wikibase, Quarry, and the Wikimedia Central Auth architecture. Jupyter Hub provides the capability of users to create research notebooks that will query the datasets and provide analysis tools in several popular languages (e.g., Python and R). Wikibase will provide the structure for actually storing the manifests. Quarry enables a web-based SQL query engine for data sets that have their repositories (data) stored and accessible through our infrastructure and tooling. The Wikimedia Central Auth architecture enables the authentication of users. All users assigned to a particular repository will have shared access to slices of the repository, Python and R scripts run against the repository and researcher notes. The Provenance and Collaboration View will show the history of ingestion of the data, as well as provide a version controlled view of related analysis scripts and notes. The OCDX Repository Engine, in this case, provides a shared, generalized instance of tools that are bricolaged together in most online community research labs today. A proof of concept for this accomplishment has already been achieved; our proposal is built on a demonstration that our team has figured out the combination of tools and practices that are leading to adoption of our approach.

All OCDX tooling and infrastructure components are agnostic with respect to how OCDX manifests are applied in practice. Collectively, the tooling and infrastructure will focus on (1) improving the speed and reliability of machine- and human-generated OCDX manifests throughout dataset lifecycles, (2) fostering the sharing and discovery of published OCDX manifests, and (3) revealing how OCDX manifests are used in the practice of science. All tooling and relevant infrastructure components will be open source projects managed through GitHub and distributed under the MIT permissive license.

5. WORKSHOP PLAN

Workshop participants will be provided advance access to the open online community data, and a Jupyter Hub/Wikibase metadata management infrastructure. Examples of candidate data sets, pending evaluation of Terms of Use, include the Yelp Dataset Challenge data, Reddit images and comments, Stack Exchange data, and FLOSSmole data. An agenda with workshop plans and expectations will also be distributed in advance. Participants will be encouraged to examine the data, consider which other participants with whom they may wish to work, and prepare questions or ideas to work on during the workshop.

The workshop will open with a brief overview of the format, goals, and plans for the day (10 minutes). Each participant will provide a brief introduction to their work, their applicable skills, and specific interest in the workshop themes (30 minutes). We will then engage a brief brainstorming process, eliciting the questions and ideas attendees have considered in advance and iterating on these ideas (30 minutes). The participants will then be able to self-select into small groups of 3-4 individuals to pursue data-driven hacking oriented toward developing and implementing one or more measures of content quality or contributor performance in the shared dataset. The workshop organizers will join in with groups as needed.

Before lunch, we will break briefly to give two-minute status updates for each group, both to acknowledge the work completed to that point and to seed meal time conversation. Following lunch, the organizers will lead a short activity for the full group intended to bring awareness to the spectrum of emerging challenges and perhaps foster a new set of connection points for participants to work together. Participants will be encouraged to change groups or to form larger clusters as appropriate. We will reserve the last 90 minutes of the session for full-group discussion during which each team will present a short debrief on their progress over the course of the day and share any datasets, visualizations, and analyses they have produced.

While the primary goal of the hackathon is to produce a functional analysis of contribution quality and contributor performance, we will endeavor to create a low-pressure environment for exploratory learning and data play. The hackathon workshop model, while loosely structured and lightly managed, was highly successful at CSCW 2014/15/16 and we anticipate it will work well for Group 2016 as well.

5.1 Potential Outcomes

There are a variety of potential outcomes from a hackathon-style workshop focused on data analysis. By creating the opportunity for participants to work together with tools and theories focused on shared data sets, we expect the workshop will create an environment suitable for professional development gains for participants: introduction to new research skills and theoretical perspectives, refinement of ideas and questions for research, and developing new collaborative relationships. New research projects and publications could also emerge from the starting point provided by co-working at the workshop. In addition, participants will gain experience with the hackathon model of peer production in a research-oriented context, which was a popular feature of prior related workshops.

5.2 Logistics

The workshop will benefit from having the following equipment available: one projector, 1-2 flip charts with markers, and snacks. We can accommodate up to 20 participants.

6. REFERENCES

[1] Bird, C., Rigby, P. C., Barr, E. T., Hamilton, D. J.,German, D. M., & Devanbu, P. 2009. The promises and perils of mining git. Proceedings from Mining Software Repositories, 2009. MSR'09. 6th IEEE International Working Conference on.

[2] Blincoe, K., Valetto, G., & Goggins, S. 2012. Leveraging Task Contexts for Managing Developers' Coordination. Proceedings from ACM Conference on Computer Supported Cooperative Work, 2012, Seattle, WA.

[3] Dabbish, L., Stuart, C., Tsay, J., & Herbsleb, J. 2012. Social Coding in Github: Transparency and Collaboration in an Open Software Repository. Proceedings from CSCW'12, Seattle, Washington.

[4] Goggins, S., & Petakovic, E. 2014. Connecting Theory to Social Technology Platforms A Framework For Measuring Influence in Context. *American Behavioral Scientist, Online First.*

[5] Goggins, S. P., Mascaro, C., & Valetto, G. 2013. Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology, 64*(3), 516-539. doi:10.1002/asi.22802

[6] Howison, J., & Crowston, K. 2014. Collaboration Through Open Superposition: A Theory of the Open Source Way. *MIS Quarterly, 38*(1).

[7] Jisc (n.d.). DCC curation lifecycle model. Retrieved from: http://www.dcc.ac.uk/resources/curation-lifecycle-model

[8] Maron, D.; Missen, C.; McNeirney, K. & Elnora, K.T. 2015. Lo-fi to hi-fi crowd cataloging: Increasing e-resource records and promoting metadata literacy within WiderNet. Poster presented at the iConference.

[9] Spinellis, D., Gousios, G., Karakoidas, V., Louridas, P., Adams, P. J., Samoladas, I., & Stamelos, I. 2009. Evaluating the Quality of Open Source Software. Electronic Notes in Theoretical Computer Science, 233, 5-28. doi:10.1016/j.entcs.2009.02.058