# The Open Community Data Exchange: Advancing Data Sharing and Discovery in Open Online Community Science

Georg J.P. Link
University of Nebraska at Omaha
Omaha, NE USA
glink@unomaha.edu

Matt Germonprez
University of Nebraska at Omaha
Omaha, NE USA
mgermonprez@unomaha.edu

Sean Goggins
University of Missouri
Columbia, MO USA
gogginss@missouri.edu

Jeff Hemsley
Syracuse University
Syracuse, NY USA
jeff.hemsley@gmail.com

Bill Rand
University of Maryland
Adelphi, MD USA
billrand@gmail.com

Megan Squire
Elon University
Elon, NC USA
msquire@elon.edu

## ABSTRACT
While online behavior creates an enormous amount of digital data that can be the basis for social science research, to date, the science has been conducted piecemeal, one internet address at a time, often without social or scholarly impact beyond the site's own stakeholders. Scientists lack the tools, methods, and practices to combine, compare, contrast and communicate about online behavior across internet addresses or over time. In response, we are building the infrastructure for computational social scientists, social scientists, and citizens to make corresponding advances in our understanding of online human interactions. In this paper, we present our effort to specify the Open Community Data Exchange (OCDX) metadata standard to describe datasets, as well as the necessary infrastructure for creating, editing, viewing, sharing, and analyzing manifests. The purpose of this paper is to communicate the current state of our project and represent our current findings through our ongoing engagement with the scientific community and to engage in dialog among computational social scientists.

## CCS Concepts
• Information systems • Document representation • Human-centered computing Collaborative and social computing • Social and professional topics • Professional topics • General and reference • Design

## Keywords
Dataset Sharing. Social Science Datasets. Science of Science. Metadata Standard Development. Tool and Infrastructure Development.

## 1. INTRODUCTION
Learning, governance, social engagement, emotional support, and other basic human needs are now woven together through a myriad of online websites. Citizen scientists share data through open online communities like ebird.org. Students use the Internet for improved learning on Khan Academy. Citizens read their news online via Reddit. Patients seek information and emotional support through internet forums, such as PatientsLikeMe, and track their health information through the instrumented self-using devices like Fitbit. Increasingly, life has an online element that is essential for daily activities while at the same time serving as a reflection of related offline behaviors.

Online behavior creates an enormous amount of digital data that can be the basis for social science research. Such behavioral data has been used for research in diverse online contexts, such as scientific advances [5], online learning outcomes [1], political use of social media [8], citizen engagement [10], group identity formation [9], and valued health benefits [6]. To date, however, this science has been conducted piecemeal, one internet address at a time, often without social or scholarly impact beyond the site's own stakeholders. Thus, there is an urgent scientific need to make sense of human behavior across technologies, and an urgent human need to better understand how to apply online technologies for social benefit. To address these scientific and human needs we propose a cyberinfrastructure that will enable researchers to effectively look across online contexts to explain, in more general terms: (1) how online interactions affect participants, groups, and society as a whole; and (2) how to design online communities and platforms to maximize their positive effects.

Addressing these issues requires the systematic sharing and analysis of datasets that are currently fragmented and unavailable to most researchers. Scientists lack the tools, methods, and practices to combine, compare, contrast and communicate about online behavior across location and over time. This is not because the differences across sites are poorly understood. Goggins [4], for example, provides a coherent ontological framework for classifying online human interactions as principally between people and each other (i.e., online health forums) or people and artifacts (i.e., ebird.org). To advance science beyond a deluge of studies focused on singular sites for online human interaction, we develop an infrastructure where scientists can systematically share, annotate, analyze, and integrate data from multiple online sources.

In biology, Genbank enables scientists to share, describe, and leverage data from hundreds of labs, accelerating the development of knowledge about the human genome. Like Genbank, we are building the infrastructure for social scientists, computational social scientists, and citizens to make corresponding advances in our understanding of online human interactions.


*OpenSym 2016*, August 17–19, 2016, Berlin, Germany.

DOI: http://dx.doi.org/10.1145/12345.67890

Specifically, the large volume of online behavioral data, combined with its poor description to date, creates a number of persistent research challenges that (1) limit the discovery and reuse of large datasets built from these traces; (2) hinder researchers in combining or comparing datasets; (3) fail to provide proper attribution for those creating the datasets; and (4) make the study of how scientists are creating and using datasets in scientific inquiry difficult. In short, scientists lack the tools, methods, and practices to combine, compare, contrast, and communicate about online behavior over time and across online locations.

Understanding how online human interactions represent and contribute to learning, governance, social engagement, emotional support, and a myriad of other social scientific constructs requires coherent metadata standards and infrastructure. The authors have constructed a prototype system that allows for the sharing and analysis of online community data on a massive scale. Scaling up the approaches and practices already developed by the authors will increase the capacity of scientists and citizens who study online human interactions to make systematic, valid, and coherent comparisons across time and internet addresses for the first time. Specifically, our research advances scientific standards, cyberinfrastructure, and scientific practice in four cohesive research tracks that (1) support the organization and discovery of datasets for data intensive research involving online human interactions; (2) builds scientific capacity through the multidisciplinary reuse and combination of discovered datasets; (3) enables cross-sectional and longitudinal analysis and comparison of large, online human interaction datasets; and (4) facilitates scientific discoveries about the scientific enterprise through the large scale analysis of the ways in which social science datasets are constructed and shared.

Our research and the resulting cyberinfrastructure will advance scientific understanding of online human interactions, and how those interactions evolve over time and across sites. Enabling this more systematic approach to storing, describing, analyzing, and communicating about online behavior will advance citizen science and interest in computing professions. Recommendations from our investigation into the process of science will have a direct impact on the decision making process of policy makers interested in scientific advancement and the administration of educational institutions with respect to research.

## 2. OPEN COMMUNITY DATA EXCHANGE

Online, behavioral data sets must be described consistently in order to be discoverable by others, compared with each other, and studied in aggregate. Core to this proposal is advancing the **O**pen **C**ommunity **D**ata E**X**change (OCDX), a metadata specification and robust infrastructure for long term sustainability. This project specifically builds on the prototyped capability of the OCDX, including a *bill of materials* for datasets (OCDX manifest) as derived from the OCDX metadata specification (Figure 1).

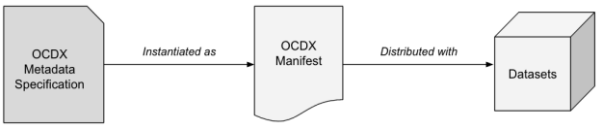


**Figure 1: The relationship between the OCDX metadata specification, the OCDX manifest, and datasets. The relationship is similar to the relationship between the W3C specification used to define HTML5 and the actual use of HTML5 in practice. The OCDX metadata specification contains details about metadata fields including acceptable formats and cardinality. The OCDX manifest is the instantiation of those details in practice.**

The precise metadata describing fundamental dataset information and recommended analytical practices are included in the OCDX manifest. The OCDX metadata standard, related OCDX manifest, and supporting OCDX cyberinfrastructure and tooling (collectively referred to as the OCDX Initiative) have been initially designed and tested by members of several scientific communities, including social science, computer science, and information systems. To date, the OCDX Initiative has been evaluated and advanced through academic and practitioner workshops in Vancouver [7], Copenhagen (May 2015), Omaha (January 2016), San Francisco (CSCW, February 2016), and Chicago (May 2016).

## 3. RESEARCH APPROACH

Technology alone will not bridge the gaps identified at the outset. Advancing scientific practices, which require people, is both more complex and more critical for success. To meet this challenge, our project will use engaged scholarship as a dominant methodological approach within which more localized methods are applied [2], as illustrated in Figure 2.

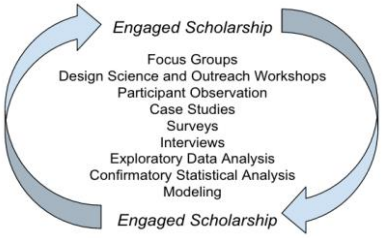


**Figure 2: Engaged scholarship as the research approach within which localized research methods are applied in the proposed project**

The pluralist approach provides context for our project and frames the setting within which we manage our project. It enables refined analyses and theoretical representation of community development, standards creation, and scientific practices that emerge as part of the OCDX Initiative [11]. For instance, within our workshops, we may conduct surveys before and after the event and interviews at the event. But, since the tools we are deploying and using in the workshops are themselves trace data collectors, we can use that data in conjunction with the surveys and interviews to create a holistic view of the experiences and events that occurred during the workshop.

To advance the OCDX initiative, a new open online community science cyberinfrastructure is designed, deployed, and managed through four integrated tracks within we will participate in engaged scholarship and our localized methods. Each track focuses on specific work in support of this goal. Additionally, research questions for each track are geared toward helping us understanding how the OCDX initiative can both improve and learn from scientific practice in the ongoing refinement of our infrastructure.

**Infrastructure Implementation -** Track 1 is aimed at creating a robust and sustainable infrastructure that supports OCDX manifest creation, governance, sharing, and access. In this effort, Track 1 advances analytic systems for the aggregation, visualization, and analysis of OCDX manifests and their use in scientific activities. We accomplish this through fostering our relationships and scaling our development efforts with the Wikimedia Foundation for robust information system platforms. Further, we will populate the information systems with an initial corpus of manifests by partnering with FLOSSmole [17] to annotate their archives, and with GitHub for continuous open online community data sourcing. Connecting the OCDX initiative with information organizations (Wikimedia), communal engagements (GitHub), and scientific endeavors (FLOSSmole) strengthens ties with our foundational,

corporate, and academic partners, fostering diverse support for the OCDX initiative. Track 1 addresses the following questions:

a. *How is massive online community data infrastructure understood, advanced, and fostered?*
b. *What are the impacts of infrastructure design decisions on the sharing and analysis of online community data?*

**Deep Dives** - Track 2 advances the integration of the ODCX infrastructure into scientific practices associated with dataset development, management, and discovery. We accomplish this by engaging with several ongoing research projects as deep-dive cases that will use the OCDX infrastructure as part of their research workflow. We will explore the ways research teams use the OCDX infrastructure in the creation of ODCX manifests. These partners include Syracuse University (political election campaigns on social media), the University of Missouri (focusing on online health support), and projects at the University of Maryland (relating online behavior to offline actions). In addition to creating a corpus of OCDX manifests generated from different types of ongoing open online community research, efforts in Track 2 will provide feedback to improve the OCDX metadata specification and supporting infrastructure. Track 2 addresses the following questions:

a. *How do formalized architectures for online community data fit within the research workflow impact the practice of science?*
b. *How can individual use cases be studied in order to gain insight to affect the development of the sharing and analysis infrastructure?*

**Outreach and Sustainability** - Track 3 is aimed at the outreach and sustainability of the OCDX initiative, requiring ongoing efforts to engage and grow the community. In Track 3, we actively connect with academic and practitioner participants through two types of OCDX-sponsored workshops recurring a total of 10 times over the course of the project. The first type of workshop includes hands-on engagement with the OCDX manifest and infrastructure as participants come to understand and advance the OCDX initiative. In this workshop, participants will integrate existing datasets with OCDX manifests and infrastructure to highlight successes and concerns. The second type of workshop will include relationship building between participants through presentations of how the OCDX initiative is currently being designed, developed, and deployed. The aim of the second workshop is to highlight real world implementations, stimulating points of common interest between participants. Both workshops are constructed with the goal of building outreach and improving sustainability of the OCDX initiative through regular and engaged community building activities. Track 3 addresses the following questions:

a. *What are key motivators for people to share their online community data and analyses?*
b. *What is the impact of outreach and sustainability efforts on promoting the sharing of such data?*

**Science of Science Research** - Track 4 is primarily aimed at advancing the science of science, with a focus on data intensive open online communities. In the fourth track, we study the scientific enterprise using OCDX manifests and infrastructure created from Tracks 1-3. In the Science of Science track, we think of the corpus of OCDX manifests as a kind of human trace data that we can study in similar ways that researchers study open online communities. We will develop analytical techniques, as well as empirical and theoretical models that leverage the OCDX manifests to help reveal the ways data intensive open online community science takes place. We will also link our findings with other published scientific data (e.g. citations) to identify factors related to scientific productivity and impact. We will demonstrate ways that the OCDX initiative will be useful in informing scientific policy associated with the systematic sharing and analysis of datasets. Feedback from Track 4 will be used to improve the OCDX metadata specification and infrastructure in ways to specifically support the science directly associated with the OCDX initiative. This is a sharp contrast to Genbank, which was designed to support sharing and discovery of data, but not to directly support the study of the scientific endeavor itself. Track 4 addresses the following questions:

a. *How does analysis of such data sharing initiatives reveal new scientific practice and inform science policy?*
b. *What is the impact of science of science findings on online community data sharing?*

The research questions in each track help us understand why and how participants engage the OCDX initiative, ways in which the OCDX metadata standard, tooling, and infrastructure are engaged, and ways that scientific metadata reveals how data intensive research takes place and becomes part of scientific practice. Figure 3 illustrates the four interrelated tracks.

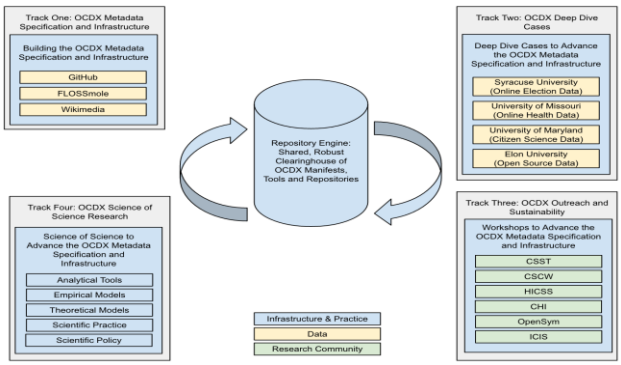


**Figure 3: Project tracks for facilitating the description and use of open online community data across scientific practice. Note that infrastructure and practice include all of the tools, metadata, repositories, hardware and scientific practice that is reflexively constructed across the four tracks.**

# 4. EXPECTED OUTCOMES

The OCDX Initiative provides metadata for researchers' intent on sharing and discovering open online community data, and studying the enterprise of open online science in hopes of informing scientific practice and policy. We are working on three primary artifacts: a metadata specification, tooling, and infrastructure. Each is introduced here.

## 4.1 Metadata Specification

The OCDX metadata specification is used to define metadata manifests to accompany partner datasets. Metadata specifications have proven valuable in bridging and connecting community members aiming to share information in the overall advancement of community health and sustainability. The Software Package Data Exchange (SPDX) community is a Linux Foundation initiative aimed at explicating license and vulnerability metadata for software packages as exchanged throughout software supply chains [3].

The OCDX metadata specification represents a key artifact from which tooling and infrastructure are derived. It is expected that through these relationships, the OCDX metadata specification will be better understood in practice, leading to its refinement to potentially include such fields as author annotations, dataset

dependencies, and dataset lifecycles. A condensed form of the current OCDX metadata specification is shown in Figure 4.

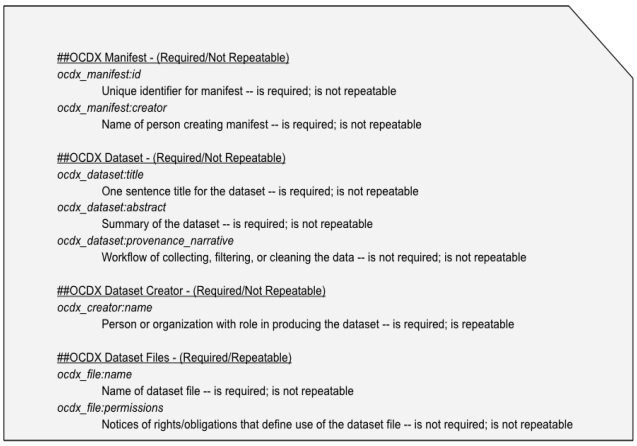


**Figure 4: The OCDX metadata specification in condensed form.**

The advancement of the OCDX metadata specification alone will move us a considerable way toward the goal of making data more reusable by a larger group of scientists. Making sure that a large percentage of open online community datasets have explicit OCDX manifests attached to them that describe what the dataset is, how it was collected, and what permissions are provided for reuse of the dataset, will make it much easier for scientists to identify datasets of interest to them, to understand datasets that were used in other contexts, and to use those datasets in their own work. Moreover, this would create labeled and related datasets demonstrating community activity, and such a set of related datasets becomes an object of study in its own right. Technology that enables the easy use of related OCDX manifests will make this work much more powerful, which is what we will describe in the next section.

## 4.2 Tooling and Infrastructure

Stemming from the metadata specification, we are advancing robust tooling through participant engaged design, development, and deployment activities. These activities involve our foundational, academic, and corporate partners. Foundationally, we are partnered with the Wikimedia Foundation to integrate OCDX tooling with existing toolkits including JupyterHub and Wikibase. Academically, we are partnered with open online community researchers to provide OCDX tooling aimed at advancing and understanding scientific practice. Corporately, we are partnered with GitHub to integrate OCDX tooling with continuously sourced community metric data. OCDX tooling includes support for the generation, management, and consumption of OCDX metadata standard derived manifests.

We propose to design and build 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 are 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.

OCDX infrastructure is aimed at supporting services by which OCDX metadata standard-based tooling is made publically available for scientific communities. The OCDX infrastructure will support public instances of all OCDX tools by which OCDX manifests are produced, managed, and discovered. Finally, the OCDX infrastructure will be available for local deployments via full source, install scripts, and documentation provided through our GitHub repository.

# 5. CONCLUSION AND FUTURE WORK

Through participant engaged design, development, and deployment, we consider the OCDX initiative as an evolving endeavor where points of interest are identified in ways that the metadata standard, tooling, and infrastructure are used, adapted, and validated. In this paper, we outlined the background and goal of this OCDX project and described our methods and outcomes. We believe that our scientific discipline will benefit from this work.

We continue to refine the specification of the OCDX metadata standard as well as the tooling and infrastructure required for open online community scientists to find, understand, create, maintain, and share dataset metadata. We invite scientists who have datasets to compile an OCDX manifest and provide feedback.

# 6. REFERENCES


[1] Bishop, J.L. and Verleger, M.A. 2013. The flipped classroom: A survey of the research. *ASEE National Conference Proceedings, Atlanta, GA* (2013).

[2] Chiasson, M., Germonprez, M. and Mathiassen, L. 2009. Pluralist action research: a review of the information systems literature*. *Information Systems Journal*. 19, 1 (Jan. 2009), 31–54.

[3] Germonprez, M., Kendall, J.E., Kendall, K.E. and Young, B. 2014. Collectivism, creativity, competition, and control in open source software development: reflections on the emergent governance of the SPDXtextregistered working group. *International Journal of Information Systems and Management*. 1, 1/2 (2014), 125–145.

[4] Goggins, S.P., Mascaro, C. and 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 (Mar. 2013), 516–539.

[5] Irwin, A. 1995. *Citizen Science: A Study of People, Expertise and Sustainable Development*. Psychology Press.

[6] Moorhead, S.A., Hazlett, D.E., Harrison, L., Carroll, J.K., Irwin, A. and Hoving, C. 2013. A New Dimension of Health Care: Systematic Review of the Uses, Benefits, and Limitations of Social Media for Health Communication. *Journal of Medical Internet Research*. 15, 4 (Apr. 2013), e85.

[7] Morgan, J.T., Halfaker, A., Taraborelli, D., Goggins, S., Hwang, T. and Computing, S. 2015. Bridging the data divide. (2015).

[8] Nahon, K. and Hemsley, J. 2014. Homophily in the Guise of Cross-Linking: Political Blogs and Content. *American Behavioral Scientist*. 58, 10 (Sep. 2014), 1294–1313.

[9] Ren, Y., Kraut, R. and Kiesler, S. 2007. Applying Common Identity and Bond Theory to Design of Online Communities. *Organization Studies*. 28, 3 (Mar. 2007), 377–408.

[10] Tandoc, E.C. 2014. Journalism is twerking? How web analytics is changing the process of gatekeeping. *New Media & Society*. (Apr. 2014), 1–17.

[11] Weick, K.E. 1989. Theory Construction as Disciplined Imagination. *The Academy of Management Review*. 14, 4 (1989), 516–531