Breaking into new Data-Spaces: Infrastructure for Open Community Science

Aaron Halfaker
ahalfaker@wikimedia.org
Jonathan Morgan
jmorgan@wikimedia.org
Yuvaraj Pandian
ypandian@wikimedia.org
Wikimedia Research

Elizabeth Thiry
Boundless
thirystheory@gmail.com

William Rand
Robert H. Smith School of Business
University of Maryland
wrand@umd.edu

Kristen Schuster
schuster.kristen@gmail.com
A.J. Million
ajmillion@gmail.com
Sean Goggins
GogginsS@missouri.edu
University of Missouri

David Laniado
Digital Humanity
Eurecat
david.laniado@gmail.com

Abstract

We propose a full day workshop focused on experimentation with documentation protocols and technology that are designed to make the process of “breaking into” a new dataset a more tractible process for researchers studying open online communities. This workshop’s purpose is to bring together researchers to test these systems, to discover problems and missed opportunities to support iteration. Participants will also be given the opportunity to use state-of-the-art documentation and technologies to break into a new collection of datasets. This workshop is the direct result of a call to action to build shared infrastructure for data sharing between researchers from past workshops at CSCW and related conferences.

Author Keywords

Open Collaboration Data Factories; Infrastructure; Methods; Policy; Online communities; Scientific practice

ACM Classification Keywords

H.3.5 [Online Information Services]: Data sharing; H.3.4 [Systems and Software]: Information networks

Introduction

Despite being easily accessible, open online community (OOC) data can be difficult to use effectively. In order to access and analyze large amounts of data, researchers must first become familiar with the meaning of data values. Then they must find a way to obtain and process the datasets to extract their desired vectors of behavior and content. This process is fraught with problems that are solved (through great difficulty) over and over again by each research team/lab that *breaks into* datasets for a new OOC. Rarely does the description of methods presented in research papers provide sufficient depth of discussion to enable straightforward replication or extension studies. Further, those without the technical skills to process large amounts of data effectively will often be prevented from even starting work. The result of these factors is a set of missed opportunities around the promise of open data to expedite scientific progress.

In this workshop, we will experiment with strategies – both technological systems and documentation strategies – designed to enable our community to more thoroughly reap the benefits of open online data science practice. We will invite participants to attempt the difficult work of breaking into a new dataset using tools and documentation designed to alleviate common difficulties. In the months leading up to the workshop, we will prepare and describe several datasets within an open querying service and invite participants to explore these systems and their functionality through the replication and extension of a selected data-intensive research paper from past CSCW. During the workshop participants will have the opportunity to explore new tools and datasets and to jump start new studies based on our curated documentation and infrastructure. As we observe and interact with our participants, we hope to learn from their successes and struggles and to use these learnings to iteratively improve our tools and documentation protocols.

This work builds on a call to action from a previous CSCW Workshops[2, 7] and ongoing initiatives[1]: to build up shared research infrastructure[11, 6] to support data and method sharing practices. The workshop organizers come from many different backgrounds and have extensive experience with using OOC data, developing infrastructure to support access to and analysis of OOC data, and building communities of practice around OOC research.

We will use this workshop to achieve three goals:

1. identify common challenges and novel strategies for making open community research easier to replicate and extend – specifically targeting protocols for documenting research methods (e.g. the ODD protocol[3])
2. inform the design of data management/analysis infrastructures like Quarry, our experimental open querying service[2]
3. inform the design of metadata indexes like the Open Collaboration Data Factory's wiki[3]

We also hope to foster a community of practice within CSCW around open data management and plan for next steps towards accelerating scientific progress around the study of computer-supported cooperation just as past workshops have informed our plans for this workshop proposal.

Using open online data

Regretfully, the technical availability of OOC datasets has not been a panacea for the study of socio-technical phenomena in these communities. Based on past research and workshops designed to help us explore OOC data science practices, we have identified three key hypotheses

[1] http://www.datafactories.org/
[2] https://meta.wikimedia.org/wiki/Research:Quarry
[3] http://wiki.urbanhogfarm.com/index.php/Main_Page

about what makes *breaking into* new datasets so difficult: (1) methods descriptions are often insufficient as a guide to replication & extension, (2) technological literacy bars access to processing large datasets, and (3) inconsistent and poorly indexed metadata prevent discovering what data is available and what the items of a dataset *mean*.

Methods replication

OOC research has advanced considerably in the last few decades, but it is still difficult to compare and contrast research findings from different pieces of work. Part of this is due to the very nature of the research; it comes from all sort of fields from information systems (e.g. [5]) to computer science (e.g. [9]) to information science (e.g. [8]) to marketing (e.g. [4]), and is studied in a wide variety of platforms from Twitter (e.g. [10]) to Wikis (e.g. [1]) to question-and-answer forums (e.g. [12]). As a result, different disciplines and different study venues use different language, and different descriptions, making it hard to integrate knowledge gleaned from different origins.

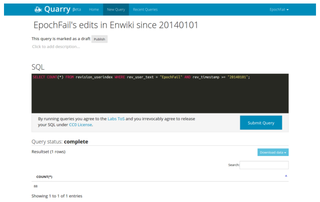
Moreover, the lack of easy translatability between fields and platforms has made the reproducibility of findings very difficult. A researcher in one field may take certain definitions for granted that are not well understood in another field, or at least are not understood in the same way. In other words, it's difficult for a researcher who works with Flickr data to understand how certain concepts are operationalized by researchers who work with blogging data. The field has progressed fine up until this point because there is so much research to do in this space, but it is now time to start to build a cohesive theory of online communities and to create knowledge built on top of other knowledge, and to provide standards that allow different researchers to reproduce each others' findings.

In order to take the field to the next step, it is necessary to develop a standard of communication that will allow different researchers to communicate how and why they performed their analysis and research the way that they did. One development in other interdisciplinary fields that has helped communication across boundaries is the creation of a uniform methods protocol (e.g. [3]). By creating such standards that describe how data is collected and analyzed, as well as how certain measurements in the theory are operationalized within the data, it is possible to make it easier for different researchers to understand each other's work and to reproduce findings.

Technical literacy

While there are many powerful, widely available, free/libre tools for gathering, manipulating, and analyzing large datasets, CSCW is an interdisciplinary field and researchers' expertise for using these tools varies quite widely. Even for researchers with such expertise, the beginning of a large-scale analysis is fraught with technical issues around formatting, types and structure and this results in a long process of trial and error. For researchers without such data engineering expertise these problems can seem intractable. For example, at a past OOC data analysis workshop that we organized for the GROUP'15 conference, our expert participants spent 5 hours (the majority of the workshop day) converting and loading (100m row) datasets into an analysis framework that would allow the larger group to answer basic questions. Even after the data was loaded, there were substantial concerns about inconsistencies between the documentations and the observed row counts.

With the goal of democratizing data analysis, we have been experimenting with open dataset interfaces that make the allow us to do such basic data engineering work up front and minimize the difficulty that future researchers experience when *breaking into* the dataset. We have identi-



EpochFail's edits in Enwiki since 20140101

This query is marked as a draft

Click to add description...

SQL

```
SELECT COUNT(*) FROM revision_userindex WHERE rev_user_text = "EpochFail" AND rev_timestamp >= "20140101";
```

By running queries you agree to the Labs ToS and you irrevocably agree to release your SQL under CC0 License.

Query status: **complete**

Resultset (1 rows)

COUNT(*)
88

Showing 1 to 1 of 1 entries

Figure 1: A screenshot of the Quarry public querying system

fied two components that characterize open dataset interfaces: (1) public GUI sandboxes and query interfaces for lightweight in-situ data exploration and (2) approachable query languages (e.g. SQL).

We have developed such an open dataset interface for Wikimedia Datasets in the form of a public SQL querying service, called Quarry. Quarry loads row-based datasets into a relational database management system and allows a user to join and filter datasets on the server through a web-based user interface. This service allows both the direct download of datasets and download/sharing of secondary queries produced by queries. We have found that non-experts can acquire proficiency in SQL over the course of an hour and that experts can use SQL powerfully. Further, by making past queries public, newcomers are able to learn common and advanced querying strategies on their dataset of interest. This helps non-experts to quickly gain proficiency and, thus, become increasingly comfortable with new technologies that support their research agendas.

We see querying interfaces like these as a key opportunity to make OOC datasets more accessible to both data engineering experts and laypeople. In this workshop, we'll put this conjecture to the test by supplying datasets through Quarry and learning from the experiences of participants.

Metadata and taxonomies

In order to *break into* a new dataset, a researcher will need to discover it and determine know how to make use of it. Currently, OOC datasets are scattered across various websites on the internet. They are inconsistently (if at all) documented and the terms used to describe the characteristics of the data differ based on the discipline of the authors. By gathering and standardizing information about OOC datasets, we can dramatically improve the discover-ability and utility of them.

Classifying OOC datasets so interdisciplinary researchers can discover, access, and use them in collaboration with other scholars requires consistent and agreed upon descriptions. Engaging these challenges in a CSCW workshop will allow us to articulate shared research goals, develop common terminology for describing datasets in generalizable terms, and determine how to document metadata at different descriptive levels so that OOC researchers can use these datasets effectively.Âă

OOC datasets can be described on three levels: the meta, mezzo and micro. The meta-level is descriptive information that aids researchers in finding datasets and conducting a preliminary evaluation of their value prior to use. This level also supports data management[4]. The mezzo-level describes the meaning captured in a dataset's content. Scholars use mezzo-level information to negotiate and analyze to understand OOCs across disciplines, create theories, conduct scholarship, etc. Last is all micro-level dataset information, which granularly describes the contents and structure of a dataset. Tiered metadata schemas allow us to account for different research methods, modes of analysis, storage systems, and disciplinary norms and to support other considerations such as dataset accessibility (e.g. copyright) and research ethics.

Workshop plan

Participation

Participants will be recruited through a mixture of strategies. We will contact participants from past workshops[11, 7, 2]. We'll post announcements on social media (Twitter & Facebook) as well as open science/HCI related listserves. Participants will be selected based on their interest and experience working on OOC datasets. We will be inclusive since we'd like to learn about opportunities to support a

[4] http://www.dcc.ac.uk/resources/curation-lifecycle-model

wide variety of research experience levels and expertise, but we will cap attendance at 50 participants because it would likely become difficult to manage after this point.

As we accept participants to the workshop we'll survey them to gather ideas for replication/extension studies that could be mostly completed in the context of a Workshop. Participants will be asked to suggest a study to replicate/extend and describe what datasets and analysis would be necessary.

Workshop preparation

We'll gather and describe a small set of primary datasets relevant to the replication/extension studies we plan to ask participants to run. We'll supplement the methods descriptions of the paper we have chosen to replicate based on our proposed methods protocol. As part of this work, we'll also perform our own replication in advance to know what time-intensive analyses are involved. We'll take the opportunity to produce secondary datasets that would take too long to reproduce in the course of an 8 hour workshop day.

Datasets will be preloaded in our shared querying environment (Quarry) and metadata will be described in the OCDF metadata census wiki. Both of these systems work as intended today, but we'll be continuing to extend them and add features as the workshop approaches.

Workshop day

Vision statement. A short presentation and extended discussion about the purpose of the workshop and the larger initiative towards better infrastructure for open community data science.

Hack session. Participants (split into teams) work on the replication/extension task. Participants will have a total of 4.5 hours total for time on task besides introduction, breaks, and reflection time. The workshop organizers will work with participants to both answer their questions and observe their work.

Reporting and reflection. Participant teams report on their progress and reflect on what did and did not work for them. We will specifically ask how the methods description, querying system, and metadata was helpful and how.

Schedule (tentative)

- 8:15-9:00: breakfast mingling
- 9:00 (sharp!): AH intro to the day (process + brief overview task)
- 9:10-10:00: Vision statement about Infrastructure for OOC studies
- 10:00-10:15: Data introduction – Each team/table reviews the task, documentation and infrastructure.
- 10:15-10:30: coffee break, email breaktime
- 10:30-12:00: Morning hack session breakouts (one team per table)
- 12:00-12:30: Lunch serving, email breaktime
- 12:30-3:15: Afternoon hack session breakouts (one team per table)
- 3:15-3:30: coffee break, email breaktime
- 3:30-4:30: Report-out and reflection
- 4:30: Wrap-up, Thanks & Next steps.
- 5:00: Dinner discussion & share contacts.

Summary reporting

At the end of the day, we will use the last hour as an opportunity for our participants to discuss what worked and what didn't. We will capture their discussion points in a collaborative document that all participants will be invited to edit and extend. We'll use these notes and our observations during the workshop to publish a report summarizing major take-aways to inform future work.

REFERENCES

1. Ivan Beschastnikh, Travis Kriplean, and David W McDonald. 2008. Wikipedian Self-Governance in Action: Motivating the Policy Lens.. In *ICWSM*.

2. Sean Goggins, Andrea Wiggins, Susan Winter, and Brian Butler. 2014. OCData Hackathon CSCW 2014: online communities data hackathon. In *CSCW*. ACM, 317–318.

3. Volker Grimm, Uta Berger, Donald L DeAngelis, J Gary Polhill, Jarl Giske, and Steven F Railsback. 2010. The ODD protocol: A review and first update. *Ecological modelling* 221, 23 (2010), 2760–2768.

4. Robert V Kozinets. 2002. The field behind the screen: Using netnography for marketing research in online communities. *Journal of marketing research* 39, 1 (2002), 61–72.

5. Meng Ma and Ritu Agarwal. 2007. Through a glass darkly: Information technology design, identity verification, and knowledge contribution in online communities. *Information systems research* 18, 1 (2007), 42–67.

6. Jonathan T Morgan, Aaron Halfaker, Dario Taraborelli, and Sean Goggins. 2015a. Advancing an Industry/Academic Partnership Model for Open Collaboration Research. https://meta.wikimedia.org/wiki/Research:Open_Collaboration_Systems_Workshop/Report. (2015). Accessed: 2015-10-10.

7. Jonathan T Morgan, Aaron Halfaker, Dario Taraborelli, Tim Hwang, and Sean Goggins. 2015b. Advancing an Industry/Academic Partnership Model for Open Collaboration Research. In *CSCW*. ACM, 293–296.

8. Jenny Preece. 2000. *Online communities: Designing usability and supporting socialbilty*. John Wiley & Sons, Inc.

9. Ben Shneiderman. 2000. Creating creativity: user interfaces for supporting innovation. *TOCHI* 7, 1 (2000), 114–138.

10. Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *CHI*. ACM, 1079–1088.

11. Andrea Wiggins, David Gurzick, Sean Goggins, and Brian Butler. 2014. Quality Hackathon: Evaluating the Products of Online Co-Production Systems. In *GROUP*. ACM, 321–323.

12. Jun Zhang, Mark S Ackerman, and Lada Adamic. 2007. Expertise networks in online communities: structure and algorithms. In *WWW*. ACM, 221–230.