Characteristics of Collaboration in the Emerging Practice of Open Data Analysis

Joohee Choi
University of Maryland
College Park, MD
jchoi27@umd.edu

Yla Tausczik
University of Maryland
College Park, MD, USA
ylatau@umd.edu

ABSTRACT

The democratization of data science and open government data initiatives are inspiring groups from civic hackers to data journalists to use data to address social issues. The analysis of open government data is expected to encourage citizens to participate in government as well as to improve transparency and efficiency in government processes. Through interviews and survey responses we gathered information on forty projects that involved the analysis of open data. We found that collaborations were interdisciplinary, small in scale, with low turnover, and synchronous communication. Most of the projects asked exploratory questions and made use of descriptive statistics and visualizations. We discuss how these findings contribute to an understanding of the emerging practice of open data analysis and to a broader understanding of open collaboration.

Author Keywords

Open data, Data analysis for social good, Coordinated action

ACM Classification Keywords

H.5.3. Group and Organization Interfaces: Organizational design

INTRODUCTION

Data science is growing at a rapid rate, fueled by greater availability of data and advances in the tools and techniques used to analyze it. In both industry and the public sector, businesses, governments, journalists, and activists increasingly pursue data-driven approaches to decision-making. Advances in technology have also helped to "democratize data science" by making data science easier and more accessible [6]. With less training and cost, individuals can now make use of data wrangling tools (e.g. Open Refine), pre-packaged machine learning algorithms, and data exploration tools (e.g. Tableau). Many of these technologies make integration of data science and the web easier, such as the storage of data on the web with cloud computing and interactive visualization with tools like D3.js or Google fusion tables.

This democratization of data science has encouraged data scientists, broadly construed, to address problems that promote social good [6]. The movement toward open data and the efforts taken by governments and public interest organizations like the World Bank have also pushed in this direction. Open government data is used by civic hackers, often volunteers with software development skills, to take a "creative and often technological approach to solving civic problems" [34]. Organizations, such as DataKind and Data Science for Social Good coordinate volunteers to work with data for good causes. With the rise of data journalism [28] and the increasing availability of open data, data journalists use data to improve government transparency and investigate social issues.

In this paper we study existing uses of open data for social good in order to gain a better understanding of this emerging area of coordinated action. We are interested in both how these projects are organized and the sorts of questions which they study. These projects involve many different types of actors: independent data scientists, civic hackers, data journalists. These actors come from many different backgrounds: software development, government, non-profit. There are many different sources of data and many pressing social problems. However, all individuals engaged in this work face the shared challenge of analyzing messy data for intangible outcomes. Through interviews and surveys we gather individuals' experiences in their own analysis of open data.

RELATED WORK

Use of Open Government Data

A growing number of governments provide data to the general public through websites and online portals [3]. In the United States, President Obama's 2009 Memorandum on Transparency and Open Government directed federal agencies to provide internal data to the general public by disclosing it online. Jetzek [15] identified two major goals for the open data movement in government: to promote democracy and to capitalize on the power of open innovation. Open government data increases transparency by allowing public oversight of government, and it encourages participation by providing a new avenue for the general public to become more involved in governance. These two aspects, transparency and participation, help to promote democracy. In addition, opening data allows the government to outsource data science to the crowd. This can efficiently make government services and technology better as well as fueling entrepreneurship and innovation [15].

CSCW '17, February 25-March 01, 2017, Portland, OR, USA

© 2017 ACM. ISBN 978-1-4503-4335-0/17/03...$15.00

DOI: http://dx.doi.org/10.1145/2998181.2998265

This new abundance of government data provides fodder for civic hacking. Civic hackers make use of open data to build software applications with the aim of providing transparency and better understanding of government functions. Moreover, there is an emerging community of non-profit organizations, startups, independent information technologists and volunteers engaged in this analysis [17]. One example is a local community group in Chicago that created an interactive visualization of all lobbyist activity in the city, including lobbyists, lobbying firms, clients, and actions sought by lobbyists from the city (ChicagoLobbyists.org). Other examples include hackathons, such as the Green Hackathon, that have brought together 20-60 individuals with broad expertise to work on societal issues and has resulted in software products that use open data [39]. One such application combined supply chain information with child labor data from the UN to provide an estimate of the likelihood that child labor was used in the manufacturing of specific products.

Several papers from the HCI and CSCW communities have described some isolated uses of open government data within specific domains. For example, researchers have studied the use of Geographic Information System (GIS) data to empower regional communities [35, 38] and tax data to engage citizens with the tradeoffs in government spending [19]. Others have studied certain practices in the area of open data. Bohner and Disalvo [2] interviewed the leaders of civic tech in Atlanta, finding that openness in government data is more a spectrum than a binary. Erete and colleagues [9] showed that non-profit organizations use data-driven stories as arguments to potential funders and stakeholders. As yet, there have been no attempts to give a broad overview of the analysis of open data from a CSCW perspective.

Data Science and Data Analysis

Advances in hardware and software technologies have led to a rapid increase in the amount of data collected. Companies and organizations are recognizing the advantages of using this data in decision making and hiring people with the skills to exploit this data. This has led to the burgeoning field of "data science". Despite the recognition of the importance of data science and the need to train data scientists, the field and skills are fuzzily defined [31]. Data scientists are expected to make meaning from data using a broad collection of skills. There is little to no academic research about data scientists and their work practices. Instead a majority of the discussion has come from position articles in popular media. Harris and colleagues [13] have argued that data scientists come from many different backgrounds that draw analytic skills from five different areas: business, machine learning, math, programming, and statistics. A perfect data scientist is often described as a 'unicorn' because it is impossible for an individual to have all the skills needed. Renowned data scientists have urged their field to make use of more teams because it is so difficult for any individual to gain a complete skillset [29, 30].

Collaboration is common in the practice of statistics, one of the parent disciplines of data science. One frequent type of collaboration is between a set of domain scientists and one or more statisticians [16]. Data analysis has multiple stages, from problem formulation to data collection through analysis to conclusion [24]. Collaboration and communication between domain scientists and statisticians is important throughout all stages, but particularly during the problem formulation period. Because the domain scientist may not clearly formulate the problem, the goal of the statistician is to listen and draw out the nature of the problem, then reformulate it in a way that can be tested statistically [18]. In this way the statistician establishes "a mapping from the client's domain to a statistical question" [12]. Chatfield [5] argues that statistical tasks are tricky because the context of the data matters: there is often messiness in the data, and the objectives of the analysis are not necessarily clear. The statistician is encouraged to ask many questions of the domain scientist to gain background information and context to understand the data. Because communication during this period is both difficult and critical, Chatfield suggests the following: "from bitter experience, I particularly advise against consulting by telephone or electronic mail, where one cannot see the data". In this type of collaboration domain scientists provide understanding of the problem, the goals, and the data; statisticians provide the technical skills to construct the appropriate analysis and extract meaningful results.

Open Science, Open Collaboration, and Open Innovation

Open data analysis shares commonalities with several forms of collaboration in which sharing and openness are important tenets. There is a movement toward more open sharing in science, particularly of data. Data sharing holds scientists accountable by allowing others to confirm findings. Data sharing also accelerates scientific progress through the reuse of a valuable resource [23]. In spite of these advantages, data sharing in science is difficult [1]. One obstacle is the willingness of scientists to share their data. There is a tradeoff between cooperation and openness on one hand and competition and secrecy on the other [36], and different scientific disciplines adopt different norms of openness. Another difficulty is in the use of shared data. Scientists must assess whether a given dataset is relevant, whether they can understand the data, and whether they trust the data before deciding whether or not to reuse data [10]. Data often lacks adequate documentation to understand the context in which it was created, its format, and the meaning of its fields [1]. Understanding the data often requires interaction with one of its creators [32]. Open data analysis, like the movement toward open science, involves the sharing and reuse of open data.

Open data analysis also involves the joint production of a shared artifact. Forte and Lampe [11] define open collaboration as online collaboration that satisfies four conditions. It must produce a shared artifact, collaboration must be supported by a technological platform, this platform must allow for contributors to enter and exit the collaboration easily and the platform must allow for flexible social structures. The two most studied, prototypical examples of open collaboration are encyclopedia editing on Wikipedia and open source software development. Easy entry into a collaboration on technologically-mediated collaboration platforms allow large-scale participation [11]. Successful open source projects can attract tens of thousands of participants [26].

However, easy exit means turnover is high in open collaboration [8]. On Wikipedia, a large majority of editors only make a few edits on one occasion [4]. While technologically-mediated communication helps to facilitate large-scale collaboration by reducing the costs of communication, it may not be well suited for collaboration that requires high levels of iterative feedback between participants [11]. Technologically mediated communication often lacks the richness needed to establish common ground and support tightly coupled work [27].

Open data analysis also shares similarities with the Do-It-Yourself (DIY) and maker movements. The maker movement is the practice of working with materials (e.g. electronics, fabrics) and fabrication tools [22]. Some have argued that it represents the "democratization of technological practice" [33]. Like work with open data, many participants embrace a hacker ethos in which creating, playfulness, and tinkering are encouraged [37]. Offline collaborations in hackerspaces are as important as online spaces [33, 22]. It has a mix of lay experts and professionals and has been described both as a hobby activity as well as a form of open innovation that leads to the creation of professional manufacturing products [22]. Wang and Kaye [37] argue that it has looser community boundaries than traditional communities of practice and describe it as a collection of practice.

RESEARCH QUESTIONS

Given the large quantities of data and the complexity and multidisciplinary nature of data analysis, collaboration is likely to play an important role in the analysis of open government data. Governments make hundreds of datasets available and the most interesting and valuable analyses often come from combining several of these in a novel way [14]. Thus, no one individual can single-handedly analyze all of the available data. Even the work involved for a single project can be substantial and may require a variety of skills and knowledge. Open data is often provided in bad formats and must be extracted, cleaned, and processed; this requires coding skills. To test hypotheses and claims requires statistical knowledge. Context is often critical to understand data, to understand how statistical models fit into research questions, and to interpret the results of these models [5]. Thus, we expect that individuals working with open data would often work in collaboration with others.

One the goals of this project was to understand how collaboration unfolds in open data analysis projects. Open data analysis shares commonalities and differences with multiple forms of collaboration in which sharing and openness are important tenets, such as open science, open collaboration (e.g. open source software), and the maker movement. To address this question, we employ the Lee and Paine [20] Model of Coordinated Action (MoCA). MoCA is a descriptive model used to understand collaborative work; it expands Johansen's 1988 time-space matrix with two dimensions of synchronicity and physical distribution to seven dimensions including synchronicity, physical distribution, scale, planned permanence, turnover, number of communities of practice, and nascence. Collaborations can be characterized along each dimension. Participants either communicate at the same time or at different times (synchronicity), communication is remote or in person (physical distance), many people participate or few people participate (scale), collaborations are short-term or long term (planned permanence), turnover is high or low. Some collaborations draw participants from many different backgrounds (number of communities of practice), and these participants have different norms, practices, expertise and tools. Work practices can be established, routine, and well understood or they may be unestablished and in development (nascence).

MoCA can be used to describe meaningful differences in work practices. For example, a collaboration in which individuals come from many different communities of practice entails a culturally diverse group with different norms, practices, tools, and languages. Members can make use of their different backgrounds by engaging in work in complementary ways, but diverse backgrounds may also cause more difficulty in working together. We use MoCA to address the specific question:

Research Question 1: How do open data analysis groups coordinate their analytic activities?

Open data analysis projects use open data to produce a tangible artifact, such as a tool or a report. A second goal of this project was to understand what types of artifacts were being created. The intent of analyzing data is to produce insights, such as identifying trends, observing anomalies, and drawing meaningful inferences. These insights can be used to reflect on government practices and to suggest changes in these practices.

There are multiple approaches for extracting insights from data. By building data processing, summarization, and visualization tools, projects make data more accessible to others. Tools provide the means for an audience to find their own insights in the data and to create their own meaning from these insights. Alternatively, authors analyze data and summarize their analyses in reports. In reports, authors present the insights they discovered, their interpretation and their conclusion to an audience. There are different types of analyses, some of which are more complex than others [21]. Exploratory analyses identify trends, correlations, or relationships in the data. These analyses can be used to generate ideas, but have not been formally evaluated. Inferential analyses evaluate whether a pattern will continue to hold for new samples. Finally predictive analyses use a set of features to predict an outcome of interest for a single person or unit. The latter two, inferential and predictive analyses, require substantially more skill to apply, but can provide more reliable conclusions. We categorized projects based on the type of artifact (tool or exploratory, inferential, or predictive analysis) they created to address the question:

Research Question 2: What type of artifacts are being produced by open data analysis projects?

METHOD

Interviews

We conducted semi-structured interviews in order to understand the current practices of open data analysis. Each participant was asked to describe one particular project that they had worked on. Interview questions asked about 1) the purpose and goals of the project (e.g. "What problem (goal) does the project address?", "What are the questions that you asked about the data?", "What story did you try to emphasize in the dataset?") and 2) whether they collaborated with others and, if so, how they collaborated with each other during the process (e.g. "Who did you work with?", "What is each participant's role?", "How did you communicate with each other during the project?"). Since the goal of this exploratory research was to gain a broad understanding of current practices, we actively recruited four different groups of individuals who work with open data: data journalists, civic hackers, public officials, and professional data scientists.

We recruited participants by sending direct email invitations to data journalists (https://jplusplus.github.io/global-directory/); sending direct messages to participants in the civic hacking organization Code for America; sending direct messages to authors of data science blogs (http://source.opennews.org/en-US); and advertising on one author's Twitter accounts. We also used snowball sampling to broaden our base of interviewees. In total, 22 people participated in the interview study, though 4 were excluded from the final analysis because they did not use data that was open. For these 4 excluded cases, the analysis results were open to the public but the dataset used for each project was created privately and was not released to public. On average the interviews lasted for around 35 minutes, with the minimum of 20 minutes and maximum 50 minutes. One participant was interviewed in-person, four by video, eleven by audio, and two by email.

The authors used iterative coding based on the fundamental idea of grounded theory [7]. Initial codes were developed by one of the authors based on relevant dimensions of collaborative work practice [20], allowing for open coding of additional dimensions that might be specifically relevant to data analysis. The authors then talked through whether codes should be added or removed and settled on a final set of codes

Survey

To expand the number of study participants and to complement interview results with quantitative responses we surveyed additional participants. The questionnaire focused on the same two aspects of open data analysis projects: the goal of the project and the way in which people worked together on each project. Each participant was asked to answer the questions for the last completed project in which they had used open government data. The survey questions were developed from the responses to the earlier interviews.

We recruited participants through the same sources listed above as well as by posting recruitment message on online forums and communities (e.g. http://reddit.com/r/opendata) and relevant Facebook Groups. 32 participants began the survey and met the inclusion criteria of completing at least one project using open government data. The only incentive given to participants to complete a lengthy survey that takes 20-30 minute to complete was being entered into a raffle to win $50 dollar Amazon gift certificate; as a result, only 22 people completed all survey questions[1].

Participants

In total, forty individuals participated in this study. These participants came from a wide range of professions, from software development to journalism to data analysis and public service. Many were students and/or researchers. Most interviewees had used open data in the context of civic hacking (7, 38%) or data journalism (7, 38%). The few exceptions were, public officials who had worked with community members who were analyzing government data (2, 11%) and professional data scientists who had been hired to use open data for specific projects (2, 11%). The largest number of survey participants used data in the context of civic hacking (9, 41%). Other survey participants included students and/or researchers (7, 31%), journalists (3, 14%), data scientists (2, 9%), and unspecified others (2, 9%). Each participant was asked to select a recent project that they had actively participated in, survey participants were explicitly asked to discuss their most recent completed project.

Three quarters of participants were male (Interviews: 76%, Survey: 77%). On average participants were in their 30s (Interviews: 35% 20-30 yrs., 47% 30-50 yrs., 18% over 50 yrs., Survey: M = 33, range 22 to 50 yrs.). We specifically recruited U.S. participants to obtain a relatively homogeneous sample. Different political climates of different countries are expected to influence the use of open data. All of the interviewees lived in United States at the time of interviews. 81% of survey participants reported that they currently lived in the United States (other participants were from South Korea and Singapore). Given the low sample size we did not exclude non-U.S. participants. Their responses were not substantively different from U.S. participants. All survey participants had enrolled in some college classes and 50% had a Master's degree or higher (we did not ask educational level for interviewees).

RESULTS

Research Question 1: How do open data analysis groups coordinate their analytic activities?

One major goal of this study was to characterize the specific nature of collaboration in the emerging practice of open data analysis. Collaboration was an important part of most open data analysis projects (89%). In this section, we use Lee and Paine's [20] Model of Coordinated Action to describe collaboration in open data analysis projects along seven dimensions: scale, planned permanence, turnover, number of communities of practice, synchronicity, physical distribution, and nascence. These dimensions were coded from responses to

[1] While 22 stayed until the end of the survey, the number of participants for each question varies slightly since we did not exclude responses from the dropouts

DIMENTION	ITEM	MEDIAN	RANGE
Scale	How many people were involved in the project?	3	1 - 40
Planned Permanence	How long did the project last? (Days)	90	2 - 1,440
DIMENTION	ITEM	MEAN	SD
Turnover	How frequently did people join the project after it was started?	2.04	1.08
	How frequently did people leave the project before it was finished	1.79	1.10
Nascence	Felt uncertain about project outcomes while working on the project	3.4	1.0
	Had to make adjustments to their plans for analysis	3.5	1.1
	Lacked important context to understand the data at the beginning	3.3	0.8
Project Openness	Openness to join the project	2.7	1.7
	Accessibility of data, code, and materials	3.4	1.5
	Accessibility of end products	3.9	1.5
DIMENTION	ITEM	COUNT	PERCENT
Communities of Practice	1+ years of training in more than two skills (inferential statistics, machine learning, software development)	14	67%
	1+ years of experience in more than two domain areas (government, journalism, activism, social services and non-profits)	16	76%
Synchronicity	In person and synchronous communication	17	74%
	Asynchronous communication	6	26%
Physical Distribution	Mostly in person or a substantial mix of in person and remote	13	57%
	Mostly remote communication	10	43%
Beneficiary	Citizens of a specific region, a specific group of citizens	8	40%
	A member of the group	7	35%
	Government	3	15%
	Specific client	2	10%
Group Formation	Online	10	42%
	Through work or friends	9	38%
	Both online and through work and friends	5	21%

Table 1. Descriptive statistics for survey responses to seven MoCA dimensions (bolded) and additional items. 5 point scales were used to measure Turnover (unipolar, "Never" to "Constantly"), Nascence (bipolar, "Strongly Disagree" to "Strongly Agree"), and Project Openness (unipolar, "Completely Open" to "Completely Closed"). The number of communities of practice were calculated based on participants ratings of their own and their group members areas of expertise. All survey data was retained, even for participants who dropped out, some sample sizes are larger than 22.

semi-structured questions (e.g. *"How did you communicate with each other during the project?"*, *"Did you work remotely or in the same place?"*) asked during the interviews and from responses to specific questions asked to survey participants (e.g. *"How frequently did other people join your project after it was started?"*). While coding interview transcripts, an additional dimension, beneficiary of the project, emerged as another important aspect of collaboration. We included it as an eighth dimension.

Scale

Scale refers to the size of a group of practice. We asked interviewees and survey participants how many people worked on their project in some capacity, such as by providing feedback, providing guidance, or conducting analyses. Both interviewees and survey participants reported working on data analysis projects in small teams (Table 1). So many projects may have been small in part due to the fact that participants found others to work with primarily from people they already knew. Several interviewees described seeking out colleagues and friends as collaborators because they knew these people had the expertise they needed to analyze the data. For example Participant 13, a data journalist who worked with super PAC donation data, contacted a colleague at an organization which he knew had experience reporting on political donations. Another interviewee, who worked as a data analyst at a governmental institution, contacted a colleague who had expertise with human resources data to help provide context to understand the data (Participant 4). Participant 6, a civic hacker, said that that they often looked for people with relevant domain expertise (e.g. labor law, health care) from within their organization when starting a project.

Survey participants were evenly split, with participants finding collaborators through personal connections, such as work or friends (38%), online through organizations like Code for America (42%), or both (21%). While most groups were small, there were a few exceptions; one of our survey participants indicated that he worked on the project in a group of 40 people.

Another factor that affected scale was the degree to which project groups were open. Some projects made their materials open, allowed anyone to join, and made their end products open; others were only partially open. Participant 7 used GitHub to make all code publicly available both during and after the project. In contrast, Participant 14, a data journalist, worked with two other journalists; they compiled data from multiple sources including open and (previously) closed data, only making their results available once they were finished. On average, projects made their end products "Mostly open". Groups were bimodal in terms of making their materials open and allowing anyone to join. Data journalists on average made their project data, code and materials open, but did not allow anyone to contribute to the project. By contrast, civic hackers were more likely to make their project data, code, and materials open and to allow anyone to join. Larger groups were formed when project materials were made available and anyone could join.

Turnover

Turnover refers to the frequency with which old members leave the group or how often new members join. In general, member turnover was rarely identified in the interviews, which is consistent with our survey findings. We asked survey participants to rate how frequently other people joined or left their projects. On average, survey participants reported that other people "Rarely" joined or left a project after it began. Many civic hacking projects (e.g. Participant 6 and 7) were almost entirely open throughout the whole life cycle of the project. However, limited resources often prevented groups from actively responding to and incorporating feedback from others during development.

Planned Permanence

It was difficult to address planned permanence as described by the MoCA framework because of the decentralized, informal nature of open data analysis. The majority of groups did not start with fixed end date for their projects. Hence, rather than ask how long projects were intended to last, we asked participants how long their projects had actually lasted. Projects lasted anywhere from 2 days to nearly four years. Most projects had a specific end goal. Most of the projects reached that goal, creating a preliminary or finalized tool or report.

Number of Communities of Practice

A community of practice is a collection of people who share norms, practices, expertise, and tools. Participants and their collaborators came from multiple communities of practice, from software development to data science to journalism to city government. In describing the members of their groups, participants described collaborators with heterogeneous backgrounds. Furthermore, people with different backgrounds played different roles within the group.

Interviewees reported that within almost all of the projects there was at least one person who acted as a domain expert and at least one person who acted as a technical expert. Domain experts provided information about the larger context of the data, including explaining what was and was not captured by the data, identifying other sources of data, and identifying interesting and meaningful questions to ask with the data. Technical experts completed most of the work and provided guidance on which analytic methods to use. They also helped shape the questions by using "quantitative thinking". Participant 8, a front-end web developer, described how their team worked with a city employee who understood regulatory frameworks in order to parse the data and focus in on the most important parts. He said "she was the domain expert. I am just a software engineer ... There were like 7 different forms to fill in to enter campaign finance data and there were tons of different ways to fill out the seven forms ... she let me know the three things that she thought were most important out of all seven forms". This helped the team focus their attention on the most interesting parts of the data. She also helped them understand what was missing: only contributions of at least $100 were recorded in the data.

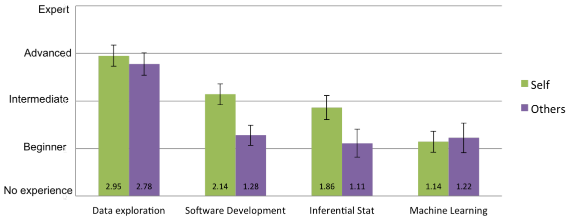

Figure 1. Mean of level of experience (standard error) estimated for each area of expertise. Participants rated their own expertise (Self) and the most expert other member of the group for each area (Others).

The role of domain and technical experts was similar to the roles of "thinker and doer" [25], where the domain experts did more of the thinking and framing of the work and the technical expert did more of the implementation and conducted the analyses.

Group members came from many different backgrounds. This resulted in groups that had a wide variety of skills and experience. Survey participants rated their own and other team members' levels of expertise (Figure 1, Figure 2). More than half of groups had group members with at least 1 year of training per subject in two or more specialized areas of inferential statistics, software development or machine learning. These are areas of skill that come from very different schools of training. Individuals skilled at software development are unlikely to be skilled at inferential statistics. Most groups also had group members with at least one year of experience in two or more domain-specific areas, such as government, journalism, activism, or non-profits. As we described in the section on scale, collaborators were often chosen specifically because they had the complementary expertise needed to understand the data.

Not only are many different communities of practice actively engaged with open government data, even within groups there are people from many different backgrounds. These people bring together a diversity of skills and practices that make such groups highly interdisciplinary. This interdisciplinarity is in part intentional, with people from different backgrounds playing different roles within the group.

Synchronicity and Physical Distribution

Many interviewees made use of regular synchronous communication. Participant 15 and her collaborator spoke on the phone regularly while they were trying to design and scope the project. Participant 15, who had more experience with data science projects, was in charge of coordinating the project. She worked with her collaborator to identify a project goal and an appropriate data set. These conversations helped them to shape the project into one that would provide tangible benefits and that could be carried out in the few months they had to work on the project. Similarly, Participant 18 met

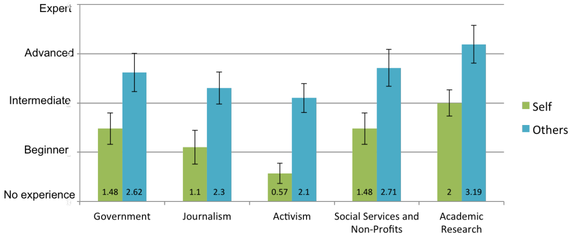

Figure 2. Mean of level of experience (standard error) estimated for each subject domain. Participants rated their own expertise (Self) and the most expert other member of the group for each area (Others).

weekly by phone with collaborators at a different site. Participant 18, a government employee, had identified a data set and a data science problem that he knew could be put into practice by one of their departments. He acted as an advisor and go-between because he understood the data and project goals. The collaborators at the other site were tasked with carrying out the analyses. These collaborators had technical skills but lacked experience with this type of data. These conversations were used to discuss ongoing analyses. Throughout the project, the collaborators used these phone calls to ask Participant 18 specific questions about the data to help orient themselves and understand the context of the data better. After making progress on their analyses, the collaborators would present Participant 18 with their progress in order to get feedback on their results. They also arranged a face-to-face meeting for all group members to meet. These conversations were valuable because they provided the technical experts with guidance, feedback, and context from a domain expert that helped them to do their work. Synchronous communication in particular was critical for some projects because of the interdisciplinary nature of the groups in which domain and technical experts took on different roles.

According to the survey, most of the groups relied on synchronous information channels such as face-to-face meetings, audio calls, and/or video calls for communication. More than half of the respondents had in-person meetings with group members. The rest of respondents answered that they mostly communicated remotely.

Nascence

Lee and Paine [20] define nascence as the degree to which a coordinated action is new and developing versus old and established. To measure the concept at an individual level, we operationalized the concept by asking the subjective assessment of uncertainty felt by an individual participator in a group. As governments have started to make more data open to the general public and easier to use [3], the analysis of this data has grown as an area of practice. Many of the communities of practice it draws from—data science, civic hacking, data journalism—are themselves new and developing. To gauge nascence from the perspective of individuals, we asked more specifically how much uncertainty they had felt when working on their projects. In particular we were interested in whether they felt that they had lacked context to understand their data or problem at the beginning, whether they had had to make adjustments to their plans for analysis, and whether they were uncertain about project outcomes while working on the project. On average, survey participants leaned toward agreement that they felt uncertainty in these areas. However, this was a very weak effect and many participants did not report uncertainty.

Differences in the level of uncertainty which individuals felt can be explained in a few ways. In retrospect, participants tended not to see as much uncertainty as they may have experienced during the process. At the beginning of the interview, Participant 18 reported that the project goals, questions, and intended outcomes had remained relatively constant from the beginning of the project. Later, when we asked the participant to walk through the stages of the project, he changed his mind and described the ways in which their thinking and goals had changed. Multiple participants indicated that uncertainty was an inherent part of data analysis. Some may have viewed this level of uncertainty as expected, a normal part of the process, while others viewed it as unexpected. Participant 10 said that making changes, especially to the questions addressed in the project, is inevitable: "Doing data analysis is way of asking questions more than answering the question. But it is not the final answer. There is always possibilities to reframe the question." One answer can lead to the discovery of another question, so there may not be a definite end to such a project. This iterative process of discovery was especially evident among data journalists. Participant 13 reported that they started out with a tentative hypothesis to test with a data set and continuously improved this hypothesis. Participant 17 also mentioned "conducting analysis until" they thought they "found newsworthy story from the data". From these results it is clear that many participants experience some degree of uncertainty in their work. It is not clear whether this uncertainty is an inherent part of data analysis or whether the uncertainty will be reduced as open data analysis practices become more established.

Beneficiary of the project

The intended beneficiary of a project emerged as another important dimension that helped to shape collaboration. This dimension is not in the MoCA framework; rather we include it in this study because it played a recurring role across multiple interviews. We define the beneficiary of the project as the intended audience and/or user of the open data analysis product(s). Many projects had an explicit beneficiary. The beneficiary might be a paying client who had come to a data scientist or civic hacker with a request for a specific tool or analysis that they planned to use (e.g. Participant 3). The beneficiary might be an organization which had a particular need or an idea of what type of tool or analysis could provide societal value (e.g. Participant 18). In these cases, the beneficiary provided strong guidance to ensure that the project succeeded in creating useful products that met their requirements. Other projects built tools and analyzed data for a general audience, often the citizens of a given region. Civic hackers spoke of building tools to inform citizens and data journalists spoke of writing articles for potential readers. In these projects, participants designed their tools and analyses with the general audience in mind. With no specific beneficiary and trying to

TYPE OF END PRODUCT	COUNT	PERCENT	EXAMPLE
Tool	14	45%	A tool for browsing political contributions in the state of Illinois
Analysis Report	17	55%	Does it take longer for people to get out of minimum wage jobs now?
Exploratory	13	42%	Are people getting more parking tickets now?
Inferential	2	6%	What is causing leakage in a manufacturing pipeline?
Predictive	2	6%	Which inspection sites are likely to violate rules in the future?

Table 2. Distribution and examples of artifacts produced by open data analysis projects. Projects coded for all 18 interview participants and 13 survey participants who optionally provided a link to their project materials.

appeal to a larger group, they had to make educated guesses about how to produce something that would effectively appeal to their intended audience.

Survey responses were somewhat consistent with interviewee responses, with the largest number of survey participants indicating that their intended audience were the citizens of a region. This was followed by projects that built tools or conducted analyses for one of the members of their group. A smaller percentage indicated that their audience was a specific client or a specific group, such as government official or agency.

Research Question 2: What type of artifacts are being produced by open data analysis projects?

To understand what types of artifacts were produced by these open data analysis projects we coded the interview transcripts and project materials when provided. Two types of projects emerged: those that conducted statistical analyses to address a specific research question and those that built a tool for end users to explore the data on their own. For projects that conducted statistical analyses, transcripts and project materials were further coded to identify the type of question: exploratory, inferential, or predictive.

Slightly under half of the projects built tools for end users (Table 2). These projects developed software programs or websites that made the data easier to use for others. Some projects built tools for readers to explore the data. In New York City, one group built an interactive map using 311 citizen complaints so that readers could explore which neighborhoods had the most rat-infested restaurants. Participant 3 helped create a visualization tool for port officials to monitor real-time international shipping price data. Using this tool, port officials could observe unexpected changes in prices that could help them detect fraud. This tool allowed end users to monitor changes in the data in near real-time. Other projects included tools to support data analysis by speeding up the processing of data (e.g. file conversion between data files). These tools empower end users to use data to come to their own conclusions. Using interactive visualizations, end users can focus in on specific data points, monitor trends over time, and make their own comparisons. The purpose of these types of projects is not to make an observation, to make an argument, or to support a decision. Instead the purpose is to make it easier for others to use the data. While some of these projects included visualization, none made use of statistical analyses.

The other half of projects aimed to extract insights from data, and these insights were often summarized in a report. The vast majority of these projects used descriptive statistics or exploratory analyses to draw insights from the data, while only a few projects used inferential statistics or predictive statistics (Table 2). Exploratory analyses focused on finding patterns in the data, such as trends over time, anomalies, or extreme values. Almost all of the exploratory projects made heavy use of visualizations. For example, Participant 12 investigated whether it takes longer to get out of minimum wage jobs now than it did in the past. For this they created a visualization of changes in the percentage of workers who held minimum wage jobs now and in the past. They then used these visualizations to make an argument that escaping minimum wage jobs does takes longer than in the past.

Predictive analyses were used to inform decisions. Participant 4 constructed projections based on census data to plan out various scenarios in planning for the future. A local government intended to place a limited number of language institutes around their county and this data project aimed to find an optimal distribution for these institutes to maximize both the number of people who would benefit as well as the diversity of immigrant communities they served.

Through interviews and surveys we found that nearly half of projects left it up to the end users to draw their own conclusion while the other half drew conclusions that relied almost exclusively on exploratory and descriptive statistics. Very few projects used sophisticated statistical analyses.

DISCUSSION

Through interviews and survey responses we gathered information on 40 projects that involved the analysis of open government data. We characterized the way in which work was coordinated and we categorized the type of artifacts produced by these projects. Three major themes emerged. One, groups were typically small, with low turnover, and relied heavily on synchronous communication. Two, interdisciplinarity played an important part in the formation of groups and the roles individuals played within these groups. Three, very few projects produced artifacts that used sophisticated statistical methods such as inferential or predictive analyses.

In these respects, open data analysis shares some similarities and differences with other forms of open collaboration. Like prototypical open collaboration (e.g. Wikipedia, open source software), the production of shared artifact was central to open data analysis; unlike prototypical open collaboration this shared artifact varied in the degree to which it was open and work on this artifact was not universally supported by a technologically mediated collaboration platform. Open data analysis projects had different levels of openness. All projects made use of data that was at least partially open and most made their end products open. However, they varied in whether the project was open to new collaborators and whether materials were open while work was taking place. Only some projects used GitHub, an online version control system with social transparency designed for software development [8]. Technologically mediated platforms with low barriers to entry and exit and flexible social structures enable the large-scale, asynchronous, high-turnover collaborations typical in most open collaboration [11]. Inconsistent norms about openness and a lack of a universal, technologically mediated platform may partially explain why we observed open data analysis collaborations which were small, with low turnover, and synchronous communication.

Open data analysis shares more similarities with less prototypical forms of open collaboration such as the maker movement and open science. Similar to the maker movement, there is no universal technologically mediated collaboration platform. In the maker movement, as in open data analysis, collaboration takes places through a variety of different means. Sharing of designs and ideas takes place in person or on a variety of online websites (e.g. Ikea Hacks website, Instructables) [33, 37]. Collaboration frequently takes place offline in hackerspaces and Fab Labs [22]. In hackerspaces individuals exchange knowledge of fabrication techniques; these spaces are used to collaborate, to learn and to teach [33]. The lack of central technologically-mediated collaboration likely shapes practice both at a community and artifact level for both the maker movement and open data analysis. Decentralization likely creates looser community boundaries; both activities are better explained as collectives of practice rather than communities of practice [37]. Decentralization also may explain why collaborations are smaller in scale.

The nature of data analysis tasks may create demands that constrain collaboration practices as well. Many projects organized work interdependently to support interdisciplinary roles within groups. Domain experts and technical experts took on the roles of thinker and doer, respectively, which required iterative feedback between these two types of experts. Using data that was collected by someone else is difficult. This is one of the challenges that scientists face in the reuse of other scientists' data. Data often lacks adequate documentation to understand the context in which it was created, its format, and its meaning [1]. Scientists often need to interact with the original creators of the data in order to fully understand it [32]. In open data analysis projects, domain experts who have more familiarity with the data play an invaluable role explaining to technical experts the meaning of data entries and fields and assessing issues of data quality. Domain experts also acted as advisors, guiding research questions and interpretation. Through back-and-forth discussions technical experts provided new results while domain experts gave feedback on these results. This pattern of feedback shares a resemblance to the back-and-forth communication between scientists and statisticians that helps statisticians turn scientific questions into statistical questions [12].

Analysis of open data requires interdisciplinary skills that a single individual rarely possesses. The task demands of interdisciplinarity engender a high level of interdependence, which in turn may explain why collaborations are typically small in scale and use synchronous communication. Many forms of technologically-mediated communication that help collaborations scale may be insufficient to support the iterative feedback required by complex, interdependent work [27].

Open data analysis is an emerging practice, in which the contributors, norms, methods, and artifacts are still developing. Currently we find that collaboration is interdisciplinary, interdependent, small in scale, with low turnover, and synchronous communication. We argue that these characteristics stem from the lack of a centralized, technologically mediated collaboration platform as well as the task demands inherent in reusing data and performing statistical analysis. We expect open data analysis as a practice to evolve rapidly. Collective norms develop over time and, while norms of openness and sharing are currently heterogeneous, they may converge towards greater openness. More openness together with the development of a technologically-mediated collaboration platform to support data analysis might facilitate the larger-scale collaborations typical of other forms of open collaboration. A greater total quantity of work can be completed with larger collaborations.

Similarly, techniques, methods, and objectives also develop over time. On average, participants were highly educated, and project groups had contributors with years of experience in relevant technical areas and subject domains. Despite these skills, research questions remained exploratory. The availability of data science technologies, which have lowered barriers to entry in data science, may not be enough to make sophisticated analyses accessible even for well-educated people [6]. For the few cases in which sophisticated analyses were used, these projects were often modeled after other existing projects. It may take time to build up a collective repository of ideas to support more complex methods and questions.

In this paper we characterized collaboration in open data analysis using the Model of Coordinated Action [20]. This paper is one of the first to apply MoCA to describe collaboration for an emerging coordinated action. This model provided a systematic framework to compare and contrast collaborative practices in open data analysis against other forms of collaboration. This paper demonstrates that MoCA is an effective framework to make task- and platform-independent comparisons. The largest challenge we faced in using MoCA is operationalization of its seven dimensions. Nascence, in particular, was difficult to measure. We chose to operationalize nascence as the degree of uncertainty individuals felt in their work. However, it was difficult to untangle whether individuals felt uncertainty because of the inherent uncertainty in discovering meaning from data or because individuals were trying to figure out which questions, methods, and tools to use in their analyses. There are also important aspects of collaboration that fall outside the scope of MoCA. For example, we found that the intended audience of the project shaped collaboration in open data projects. Future work, will be required to determine whether seven dimensions are sufficient to characterize coordinated actions.

Limitations and Future Work

The greatest limitation of this study is the low survey sample size. We gathered survey data to provide quantitative data to complement results from our interviews and to increase our sample size. Even so, we were only able to recruit a small number of survey participants, despite multiple strategies for recruiting a larger survey sample including posting recruitment messages in multiple online locales, sending personalized email messages, and providing a monetary incentive (albeit a low one). One of the challenges in studying open data analysis is that it has not yet developed a unified community of practice. This creates two complications. First, the lack of a unified community led to difficulties in recruiting a representative cross section of participants. Second, the lack of a centralized community made it hard to identify a sizable sample of the community. The participants that we were able to recruit are likely to be more actively involved in the projects than typical individuals and more likely to identify with open data as a community of practice. As a result of the low sample size, and the heterogeneity within this community, we do not believe these participants are necessarily representative of all individuals who work with open government data. Instead what the data does provide is a collection of over 40 example projects. Using this set of projects we have identified a number of patterns and themes in the way that these groups collaborate. Though these themes may not hold true for all projects, they are at least important considerations for many such projects. As an exploratory study this study lays the groundwork for future work, which will hopefully complement these findings using a broader and more representative sample.

Future work should look at open data analysis using a global sample. We specifically focused on the practices of open data analysis in a single country because different countries have very different political climates. Collaboration and the use of open data to fight government corruption in countries with substantial political repression or retribution may be very different from the forms of collaboration in the U.S.

In this paper we observed that large-scale collaborations are less typical of open data analysis than other, more prototypical forms of open collaboration. In part, this can be explained by the lack of a centralized, technologically mediated collaboration platform. Future work should evaluate this claim as well as investigate what sorts of platforms could best support data analysis. We found that some projects used GitHub, but this platform may not be well suited for data analysis. In particular, it lacks some technical capabilities such as the ability to store large quantities of data, to develop documentation and metadata for data sets, and version control that supports data cleaning and processing. We also argue that this work requires interdisciplinarity and interdependent work which may not be supported by the limited communication channels built into platforms like GitHub.

CONCLUSION

The democratization of data science and open government data initiatives have inspired groups from civic hackers to data journalists to use data to address social issues. The analysis of open government data is expected to encourage citizens to participate in government as well as to improve transparency and efficiency in government processes. We found that interdisciplinarity was important and that groups were typically small, with low turnover and relied heavily on synchronous communication. We found that most of the projects analyzing government data asked exploratory questions and made use of descriptive statistics and visualizations rather than more sophisticated questions and approaches. The emerging practice of open data analysis faces many challenges going forward, including how to tackle more complex questions, how to collaborate effectively with so many different communities of practice, and how to collaborate in ways that scale when interdependent teamwork is so important.

ACKNOWLEDGMENTS

Omitted for review.

REFERENCES

1. Jeremy P. Birnholtz and Matthew J. Bietz. 2003. Data at Work: Supporting Sharing in Science and Engineering. In *Proceedings of the SIGGROUP Conference on Supporting Group Work (GROUP'03)*. ACM, New York, NY, USA, 339–348. DOI: **http://dx.doi.org/10.1145/958160.958215**

2. Kirsten Boehner and Carl DiSalvo. 2016. Data, Design and Civics: An Exploratory Study of Civic Tech. In *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, New York, NY, USA, 2970–2981. DOI: **http://dx.doi.org/10.1145/2858036.2858326**

3. Morgan Brazillian, Andrew Rice, Juliana Rotich, Mark Howells, Joseph Decarolis, Cameron Brooks, Florian Bauer, and Michael Liebreich. 2012. Open Source Software and Crowdsourcing for Energy Analysis. *Energy Policy* 49 (2012), 149–153. DOI: **http://dx.doi.org/10.1016/j.enpol.2012.06.032**

4. Brian Butler, Elisabeth Joyce, and Jacqueline Pike. 2008. Don't Look Now, But We've Created a Bureaucracy : The Nature and Roles of Policies and Rules in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'08)*. ACM, New York, NY, USA, 1101–1110. DOI: **http://dx.doi.org/10.1145/1357054.1357227**

5. Chris Chatfield. 2002. Confessions of a pragmatic statistician. *Journal of the Royal Statistical Society Series D: The Statistician* 51, 1 (2002), 1–20. DOI: **http://dx.doi.org/10.1111/1467-9884.00294**

6. Sophie Chou, William Li, and Ramesh Sridharan. 2014. Democratizing Data Science: Effecting positive social change with data science. *In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) at Bloomberg* (2014). DOI: **http://dx.doi.org/10.1.1.478.3295**

7. Juliet M. Corbin and Anselm Strauss. 1990. Grounded Theory Research: Procedures, Canons, and Evaluative

Criteria. *Qualitative Sociology* 13, 1 (1990), 3–21. DOI:**http://dx.doi.org/10.1007/BF00988593**

8. Laura Dabbish, Rosta Farzan, Robert Kraut, and Tom Postmes. 2012. Fresh Faces in the Crowd: Turnover, Identity, and Commitment in Online Groups. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'12)*. ACM, New York, NY, USA, 245–248. DOI: **http://dx.doi.org/10.1145/2145204.2145243**
9. Sheena Erete, Emily Ryou, Geoff Smith, Khristina Fassett, and Sarah Duda. 2016. Storytelling with Data : Examining the Use of Data by. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'16)*. ACM, New York, NY, USA, 1273–1283.
10. Ixchel M. Faniel and Trond E. Jacobsen. 2010. Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work* 19, 3-4 (2010), 355–375. DOI: **http://dx.doi.org/10.1007/s10606-010-9117-8**
11. Andrea Forte and Cliff Lampe. 2013. Defining, Understanding and Supporting Open Collaboration: Lessons from the Literature. *American Behavioral Scientist* 57, 5 (2013), 535–547. DOI: **http://dx.doi.org/10.1177/0002764212469362**
12. David J. Hand. 1994. Deconstructing Statistical Questions. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 157, 3 (1994), 317–356. **http://www.jstor.org/stable/2983526**
13. Harlan Harris, Sean Murphy, and Marck Vaisman. 2013. *Analyzing the analyzers*. O'Reilly Media.
14. Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. 2012. Benefits, adoption barriers and myths of open data and open government. *Information Systems Management* 29 (2012), 258–268.
15. Thorhildur Jetzek, Michel Avital, and Niels Bjorn-Andersen. 2014. Data-Driven Innovation through Open Government Data. *Journal of Theoretical and Applied Electronic Commerce Research* 9, 2 (2014), 15–16. DOI:**http://dx.doi.org/10.4067/S0718-18762014000200008**
16. Brian L. Joiner. 2010. Statistical consulting. In *Encyclopedia of Statistical Sciences*. John Wiley {&} Sons, Inc., 1–9. DOI:**http://dx.doi.org/10.1002/0471667196.ess0409.pub3**
17. Maxat Kassen. 2013. A promising phenomenon of open data: A case study of the Chicago open data project. *Government Information Quarterly* 30, 4 (2013), 508–513. DOI: **http://dx.doi.org/10.1016/j.giq.2013.05.012**
18. Ron S. Kenett. 2015. Statistics : A Life Cycle View. *Quality Engineering* 27, 1 (2015), 111–121. DOI: **http://dx.doi.org/10.1080/08982112.2015.968054**
19. Namwook. Kim and Juho Kim. 2015. BudgetMap : Issue-Driven Navigation for a Government Budget. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'15)*. ACM, New York, NY, USA, 1097–1102.
20. Charlotte P. Lee and Drew Paine. 2015. From The Matrix to a Model of Coordinated Action (MoCA): A Conceptual Framework of and for CSCW. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'15)*. ACM, New York, NY, USA, 179–194.
21. Jeffery Leek and Roger D. Peng. 2015. What is the Question? *Science* 347, 6228 (2015), 1314–1315.
22. Silvia Lindtner, Garnet D. Hertz, and Paul Dourish. 2014. Emerging sites of HCI innovation: Hackerspaces, Hardware Startups & Incubators. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'14)*. ACM, New York, NY, USA, 439–448. DOI: **http://dx.doi.org/10.1145/2556288.2557132**
23. Karen Seashore Louis, Lisa M. Jones, and Eric G. Campbell. 2002. Sharing in Science. *American Scientist* 90, 4 (2002), 304–307.
24. Jock R. MacKay and Wayne R. Oldford. 2000. Scientific Method, Statistical Method and the Speed of Light. *Statist. Sci.* 15, 3 (2000), 254–278. **http://www.jstor.org/stable/2676665.**
25. Henry Mintzberg. 1994. The Fall and Rise of Strategic Planning. *Harvard Business Review* 72, 1 (1994), 107–114. **https://hbr.org/1994/01/the-fall-and-rise-of-strategic-planning**
26. Jae Yun Moon and Lee Sproull. 2000. Essence of distributed work: The case of the Linux kernel. *First Monday* 5, 11 (2000). DOI: **http://dx.doi.org/10.5210/fm.v0i0.1479**
27. Gary Olson and Judith Olson. 2000. Distance Matters. *Human-Computer Interaction* 15, 2 (2000), 139–178. DOI: **http://dx.doi.org/10.1207/S15327051HCI1523_4**
28. Sylvain Parasie and Eric Dagiral. 2013. Data-driven Journalism and the Public Good: Computer-assisted-reporters and "Programmer-journalists" in Chicago. *New Media & Society* 15, 6 (2013), 853–871. DOI: **http://dx.doi.org/10.1177/1461444812463345**
29. DJ Patil. 2011. Building data science teams. (2011). **http://radar.oreilly.com/2011/09/building-data-science-teams.html**
30. Gregory Piatesky. 2013. Unicorn Data Scientists vs Data Science Teams. (2013). **http://www.kdnuggets.com/2013/12/unicorn-data-scientists-vs-data-science**

31. Foster Provost and Tom Fawcett. 2013. Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Data Science and Big Data* 1, 1 (2013), 51–59. DOI:**http://dx.doi.org/10.1089/big.2013.1508**

32. Betsy Rolland and Charlotte P. Lee. 2013. Beyond trust and reliability: reusing data in collaborative cancer epidemiology research. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'13)*. ACM, New York, NY, USA, 435–444. DOI: **http://dx.doi.org/10.1145/2441776.2441826**

33. Joshua G. Tanenbaum, Amanda M. Williams, Audrey Desjardins, and Karen Tanenbaum. 2013. Democratizing technology: pleasure, utility and expressiveness in DIY and maker practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*. ACM, New York, NY, USA, 2603–2612. DOI:**http://dx.doi.org/10.1145/2470654.2481360**

34. Joshua Tauberer. 2014. *Open Government Data: The Book* (second edi ed.). Self published. **https://opengovdata.io/**

35. Alex S. Taylor, Siân Lindley, Tim Regan, and David Sweeney. 2015. Data-in-Place: Thinking through the Relations Between Data and Community. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'15)*. ACM, New York, NY, USA, 2863–2872.

36. Theresa Velden. 2013. Explaining Field Differences in Openness and Sharing in Scientific Communities. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'13)*. ACM, New York, NY, USA, 445–457. DOI: **http://dx.doi.org/10.1145/2441776.2441827**

37. Tricia Wang and Joseph Jofish Kaye. 2011. Inventive Leisure Practices: Understanding Hacking Communities as Sites of Sharing and Innovation. *Extended Abstracts on Human Factors in Computing Systems - CHI EA '11* (2011), 263–272. DOI: **http://dx.doi.org/10.1145/1979742.1979615**

38. Gemma Webster, David E. Beel, Chris Mellish, Claire D. Wallace, and Jeff Pan. 2015. CURIOS : Connecting Community Heritage through Linked Data. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'15)*. ACM, New York, NY, USA, 639–648.

39. Jorge L. Zapico, Daniel Pargman, Hannes Ebner, and Elina Eriksson. 2013. Hacking sustainability : Broadening participation through Green Hackathons. In *Fourth International Symposium on End-User Development*. IT University of Copenhagen, Denmark.