Building an Apparatus:
Reflective and Diffractive Readings of Trace Data

Carsten Østerlund, Kevin Crowston and Corey Jackson

Syracuse University School of Information Studies

Hinds Hall, Syracuse NY 13244

costerlu@syr.edu, crowston@syr.edu, cjacks04@syr.edu

Draft of 13 April 2017

Under review; please ask before citing or distributing

Building an Apparatus:
Reflective and Diffractive Readings of Trace Data

Abstract

When people interact via information systems, data are captured by the systems about the interaction as a side effect. These data, referred to as trace data, are increasingly available and interesting for research. In a sense, these systems form a kind of research apparatus, and like all advances in instrumentation, open new areas of study with great potential for discovery.

At first glance, such "big data" would seem to be most suitable for a quantitative and positivist research approach. However, we argue that considering such systems from a socio-material perspective offers insight into the challenges in analyzing trace data and suggests a rethinking of both quantitative and qualitative approaches to trace data. Building on Haraway and Barad's distinction between reflective and diffractive methodologies, we discuss and illustrate the strengths and weaknesses of these two approaches. A reflective methodology considers trace data as reflecting pre-given objects, people and practices. A diffractive methodology traces the ripples emerging from interference between trace data, information systems, users, designers and researchers.

Drawing on longitudinal study of citizen science practices that uses trace data, the paper illustrates the consequences of each methodology. We show that a diffractive methodology allows researchers to account for not only the socio-material dynamics of digital trace data but also the temporal dimension of online practices, directing the researcher's attention to how the apparatus configures and reconfigures not only trace data, but also the information system, users, system designers, and researchers.

Building an Apparatus: Reflective and Diffractive Readings of Trace Data

1. Introduction

As people increasingly interact via information systems, the data captured by such systems becomes available and interesting for research. Data captured as a side effect of system use—what we refer to as trace data—offer great potential for insight into the actual behaviours of users of these systems. Researchers usually face a trade-off between the number of subjects studied and the volume of data collected about each subject, but trace data potentially provides the ability to see every action taken by every individual user at a fine level of detail. These systems form a new kind of research apparatus, and like all advances in instrumentation, open new areas of study with vast potential for discovery.

However, consideration of the nature of trace data raises methodological concerns, leading to the question we address in this paper: what is the appropriate methodology for analyzing trace data? We draw on the literature on socio-materiality to critique quantitative and positivist approaches to analyzing trace data and to develop a socio-materially informed methodological perspective. We illustrate the concerns and the approach with examples from a long-term study of online citizen science projects in the Zooniverse.

1.1 Background: Trace data

Howison et al. (2011) define digital trace data as "records of activity (trace data) undertaken through an online information system (thus digital). A trace is a mark left as a sign of passage; it is recorded evidence that something has occurred in the past." For example, many studies have used posts on discussion fora as data; these posts are trace data of participation. Howison et al. (2011) identify three characteristics of trace data that set them apart from the kinds of data often used in information systems research, such as survey responses: "1) it is found data (rather than produced for research), 2) it is event-based data (rather than summary

data) and 3) as events occur over a period of time, it is longitudinal data". Though not included in Howison et al.'s (2011) definition, a further characteristic is that trace data are typically semi-structured, with a number of structured metadata fields (e.g., for a post in a discussion forum, the date and time, the ID of the poster, the name of the forum, possibly a previous message being replied to, ratings by other readers, etc.) and possibly additional unstructured data (e.g., the subject or content of the post).

These first two properties (found data and event-based data) taken together mean that the data are not intended as measures of a concept of theoretical interest, such as a user's attitudes or beliefs (e.g., as measured with survey items), but are rather records of actual behaviours on the system that must be interpreted to make a conceptual connection. The third characteristic, trace data being longitudinal data, raises questions about the temporality of both the data and the phenomena of interest. Evidence from events spread over time must be aggregated to describe the system at one point in time. Such aggregation can be problematic if the constructs of interest evolve or change over time.

The second part of Howison et al.'s (2011) definition of digital trace data is that the data are both produced through and stored by an information system. Trace data can be produced through direct observation, as in traditional ethnographic research that records events in a work environment. Online interaction has led to an increase in the use of data captured about these interactions. But to correctly interpret such digital trace data requires a deep understanding about the details of the specific system technology that captured the data, i.e., the apparatus. Of course, any study requires an apparatus for collecting and managing data. But researchers employing interviews or observation (for example) may take the apparatus for granted, being familiar with those approaches and the challenges they present. In this way, trace data bring to the forefront issues about the socio-materiality of the apparatus for data analysis.

1.2 Background: Socio-material critiques

Going back to Marx and the Tavistock studies, scholars have gathered and analyzed traces of organizational practices in ways suggesting that technologies, people and discourses come together in dynamic and reciprocal assemblages (Gaskin et al. 2014). The recent socio-material turn shines a bright light on this relationship by insisting that the material and social are inseparable in organizational action. The two are constitutively entangled (Barad 2003; Orlikowski et al. 2008). Boundaries between humans and the material blur. Socio-materiality highlights the nexus of doings, materialities, and discourses that people carefully enact to support certain practices (Law 2004; Suchman 2007), and offers an analytical gaze under which neither artefacts, people, nor practices can stand naked and alone, revealing inherent properties. Instead these are bound into one entity, where only concrete interactions occur between artefacts, people, and practices.

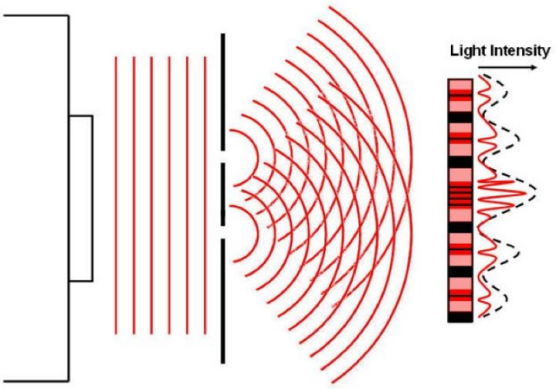
This perspective goes beyond people's mere doings. A socio-material lens highlights the performative character of action in which objects are constituted, bodies shaped, words formed, and things described. It leaves us with the question we address in this paper: how do study trace data in a way that takes their socio-material nature into account? The current IS literature on socio-materiality offers few suggestions, as the debate largely focuses on its epistemological and ontological foundation (Cecez-Kecmanovic et al 2014; Orlikowski & Scott 2013).

To address the question, we draw on two methodological metaphors introduced by Hawaway (1997) and extended by Barad (2007): *reflection* and *diffraction*. Both are optical phenomena, but the way each makes light available to an observer differs radically. A *reflection* is a representation of an object produced by a mirror. A perfect mirror produces a copy that is homologous to the original, free of distortion; a less perfect one, an image with some alterations. But in both cases, it is the reflection's likeness to the substance reflected that matters, not the nature of light producing the reflection. Observing substances from afar in the mirror, we assume that these substances are pre-given, with clear and predefined boundaries.

Both positivist and many critical and interpretivist scholars operate with a reflective methodology. Scholars with a more positivist orientation strive to accurately reflect physical reality in their data while critical scholars argue that knowledge is more accurately understood as reflections of culture. But reflection in both cases holds the world at a distance (Barad 2007). To put it differently, a reflective approach supports what Cecez-Kecmanovic (2016) describes as a substantialist metaphysics concerned with "what there is." If one assumes the primary unit of reality are self-contained and bounded substances, then one will adopt a reflective approach in the methodologies chosen to describe the properties and qualities of such entities.

In contrast, *diffraction* concerns the bending and spreading of waves when they combine or meet an obstacle. Water, light and sound all exhibit diffraction under the right circumstances. A classic example of diffraction in physics is shown in Figure 1. In this experimental set up, light from a source on the left of the figure passes through two slits in the barrier in the middle of the figure and the beams of light from the two slits interfere with each other, leaving a diffraction pattern of light and dark on the screen beyond the slits to the right of the figure. This pattern does not appear if the light shines directly on the screen or if there is only one slit. Thus, the diffraction pattern records not only differences in the source waves but their history of interactions and interferences along the way to the screen. The metaphor offers a process perspective concerned with "what is occurring" and "ways of occurring" (Cecez-Kecmanovic 2016). The primary unit of interest is not an image reflected on to a screen but the processes that created the image.

Figure 1. Diffraction pattern of light from a two-slit experimental setup.



The apparatus takes on a central position in a diffractive methodology. Barad argues that one cannot disentangle a phenomenon and the apparatus that records it. Instead, the apparatus plays a constitutive role in the production of the phenomenon by enacting specific boundaries in our socio-material reality. That is, online systems do not simply record traces of human actions and interactions, but rather actively shape them. The apparatus is not an inscription device installed before the action happens. It is not a neutral probe, measuring pre-existing entities, mere reflections of a self-contained reality. Instead, the apparatus stands out as an open-ended practice constantly producing and reproducing the phenomenon that it records.

As a result, diffractive methodologies offer a new analytical approach, in which one reads elements of the research setup *through* one another. This reading *through* is possible because the elements are intertwined: changing the size, number or position of the slits or the nature of the light source in Figure 1 causes the diffraction pattern to take on a new shape. A diffractive apparatus thus allows researchers to learn about the nature of the light source and the nature of the apparatus the light passes encounters (e.g., the slits) through study of changes in the observed pattern. For example, physicists can study the nature of a chemical element by sending light from that element through a diffraction grating with known properties and observing the resulting diffraction pattern. Reading through can also work in the reverse

direction: physicists can study the diffraction grating itself by illuminating it with light with known properties. For instance, one can learn about a crystal, used as a diffraction grating, by sending an x-ray of a known wavelength through it and studying the resulting diffraction pattern. Following the same line of thinking, information systems researchers can learn about trace data *through* studying the users of an online system, or learn about users *through* studying their information system, or learn about an information system *through* studying its traces.

Further, the practices of an apparatus are open to rearrangements. The creativity of scientific practices includes the skill of making the apparatus work for specific purposes. Elements are reworked and adjusted, leading to adjustments of the boundaries and cuts performed by the apparatus and so the nature of the phenomenon enacted and recorded. An apparatus can itself become the phenomenon, the focus of attention. This shift can happen as researchers turn their attention to the boundaries performed or by engaging the process in which the apparatus intra-acts with other apparatuses. These relations are only locally-stabilized phenomena that are part of specific performances.

In short, a focus on the apparatus as producing more than a reflection of reality allows us to 1) take the process of knowledge production into account and 2) understand our entangled socio-material world from within by reading insights through one another. Knowledge about the information system informs how we understand the traces and insights about the users help inform our reading of the traces. Moving through these different readings strengthen the overall study.

2. Case example: Setting and data gathering

To illustrate different approaches to analyzing trace data from online systems suggested by the above analysis, we present examples drawing from our ongoing research on online citizen science projects. Understanding how volunteers interact with and in citizen science projects is interesting because the activities these project support have the potential to lead to

significant scientific discoveries and to also support authentic volunteer learning opportunities (Brossard, Lewenstein & Bonney, 2005; Bonney et al., 2009; Edelson & Reiser, 2006; Wiggins & Crowston, 2011).

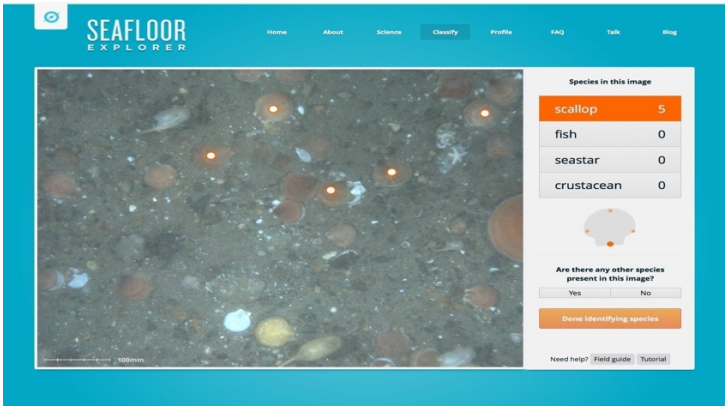
2.1 Setting: Citizen Science in the Zooniverse

Our study of citizen science draws on several years of engagement with the broader citizen science community across multiple projects (e.g., eBird and FoldIt). The specific context of our work with online system trace data has been with citizen science projects at Zooniverse. Zooniverse grew out of the Galaxy Zoo project, the prototypical example of a virtual citizen science project. Zooniverse volunteers use web-based tools to annotate image or sound observations (referred to as objects) drawn from large data sets to support scientific analysis of the objects in the data sets. The Zooniverse hosts many online citizen science projects, on a range of topics: determining galaxy morphology from galaxy images taken by the Sloan Digital Sky Survey (the original Galaxy Zoo), locating craters in images from NASA's Lunar Reconnaissance Orbiter (Moon Zoo), identifying the species of animals in photographs captured by camera traps (Snapshot Serengeti) or transcribing weather reports from scans of Royal Navy ship logs from the time of World War I (Old Weather). In each case, the scientists faced more data than they could analyze themselves, and turned to volunteers for help.

As specific example of the functionality provided by these online platforms, consider the Zooniverse Seafloor Explorer project, whose scientific goal is to better understand the distribution patterns of species in and the overall species ecology of the continental shelf off the Northeastern coast of the United States. Participation in the system starts with the presentation of a brief tutorial explaining the system and the scientific task. Many projects also provide a field guide, a digital reference assembled by the science team with exemplary images of the phenomena of interest.

After the tutorial, Seafloor Explorer volunteers view a series of photographs taken of the seafloor and for each note the presence or absence of four species of animal—sea scallops, sea stars, fish, and crustaceans—and the type of ground cover—sand, shell, gravel, cobble, boulder. The interface is shown in Figure 2, with the photograph on the left and the questions on the right. The data about an image provided by volunteers are referred to as annotations. If a volunteer identifies the presence of animals, a measuring tool appears with which to measure the size of the animal. The measurements and annotations provided by individual volunteers are then compiled (e.g., taking the consensus of multiple volunteers' annotations of an image to increase the reliability of the data) and analyzed by marine biologists to explore their scientific questions and hypotheses.

Figure 2. Seafloor Explorer annotation interface



In addition to contributing annotations and measurement data, volunteers can also engage in discussions with other volunteers and with members of the project science team on the Seafloor Explorer discussion forum (called "talk"). Volunteers will sometimes conduct their own analyses of the data objects and share them on talk. As one example, volunteers and members of the science team discussed a potential new species of underwater worm found in the images, named by volunteers the "convict worm" for its stripped pattern.

Our studies of the Zooniverse have combine engaged scholarship (Van de Ven, 2007) and virtual ethnography (Hine, 2000). The trace, interview, participant observation, surveys and

experimental data involved in this research was conducted through an engaged collaboration with developers, designers and educators at Zooniverse. Virtual Ethnography allows us to emphasize participation in the online environment through participant observation, analysis of discussion forums and project documentation and interviews. An important source of data was participant observation. As participant observers, we signed up for accounts, completed all tutorials, reviewed help resources that newcomers are prompted to review, participated in the task of classifying data and interacted in the social spaces of a project. Finally, with the cooperation of the Zooniverse developers, we had access to trace data recording the details of volunteer interactions with the system and each other.

3. Building an Apparatus: Reflection and Diffraction

As noted in the introduction, when people interact via information systems, data are captured by systems as a side effect of interaction. In the case of Zooniverse, captured data include at a minimum the annotations provided by the volunteers and postings to talk. More recently, the Zooniverse platform has been instrumented to collect essentially every action taken by volunteers as they interact with the system (e.g., reading tutorials, opening the field guide and so on). As a result, there is a large volume of trace data that can be analyzed to answer research questions about participation in such systems.

Inspired by the metaphors of reflection and diffraction, in this section we outline a series of methodologies to analyze trace data in Zooniverse. Starting with the notion of reflection, we first consider quantitative and qualitative approaches to trace data (sections 3.1 and 3.2). Turning to a diffractive reading of trace data, we then consider the building of an apparatus (section 3.3). We explore how a diffractive reading allows us to analyze processes (section 3.4) and becoming (section 3.5). Third, diffractive analysis suggests the reading of insights through one another. This approach allows us to integrate quantitate and qualitative techniques into our

analysis while maintaining a socio-material focus on processes as opposed to pre-given substances.

3.1 Trace Data Reflecting Work in the Zooniverse: Quantitative Approach

The most straightforward approach to trace data research assumes that the data *reflect* volunteer actions. Researchers explicitly (or more likely implicitly) consider the system as a source of reliable and valid data about volunteers that can be mined for patterns in volunteer behaviour. In Barad's terminology, the system is viewed as creating data that provides a reflection of the volunteers' behaviours, simply showing the researcher what happened. For example, in a citizen science project, the submitted annotations are the work submitted by different volunteers. By summing the total number of annotations associated with different volunteer user IDs, we can identify high and low contributors and look for factors associated with different contribution levels or increases in contributions from volunteers over time.

Even taking this view, trace data offer some distinctive challenges for analysis. For example, trace data often give the researcher access to a complete reflection of volunteer activity rather than a sample. However, quantitative analysts are accustomed to treating observed behaviours as being typical of usual behaviour, that is, making inferences about typical behaviours from a sample of behaviours. For example, when observing one person interacting with another in a sample of interaction, analysis often assume that that interaction is representative of a pattern of similar interactions. With a complete data set though, there is no need for that inference: the individuals either do or do not interact further. Having a complete record of a person's action means inference is unneeded and perhaps misleading (e.g., inferring a pattern of interaction from a one-off event).

3.2 Trace Data Reflecting Work in the Zooniverse: Qualitative Approach

While many quantitative approaches to trace data approach them as unproblematic reflections of user actions, qualitative researchers tend to see language or discourses as a

medium transmitting user behaviors. The data may reflect a user's actions, but to understand what they mean requires an interpretive step. In other words, one can approach trace data as a text in need of systematic analysis.

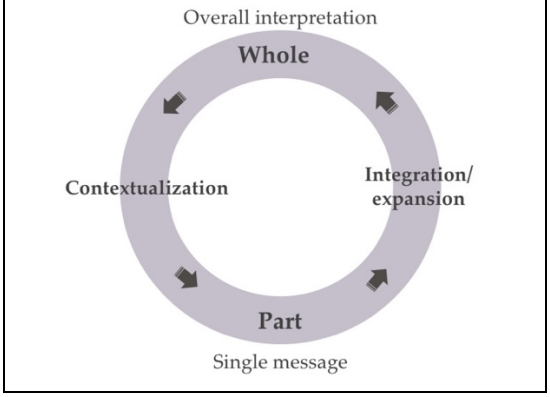
Hermeneutics offers a well-articulated approach and it has long served as a trusted pillar of qualitative and interpretive IS research. Boland (1985)—inspired by Edmund Husserl's phenomenological perspective and Gadamer's work on hermeneutics (Gadamer 1975)—was among the first scholars to introduce hermeneutics to IS research. In classic hermeneutics, the text constitutes an object of study, which is to be understood based on its own frame of reference (Kvale & Brinkmann 2009). The interpretation, for its part, aims to bring to light an underlying coherence or sense to an otherwise incomplete, cloudy, or contradictory text (Myers 1995).

More recently information systems scholars have been inspired by Ricoeur (1981) and his use of hermeneutics to study practice and social actions. Here, the researcher's position in relation to the material plays a central role in the investigation by considering the researcher's position and history regarding the study subject. Such a pragmatic and critical approach to hermeneutics allow scholars to overcome many of the shortcomings associated with a purely interpretive approach by paying attention to: 1) the broader context that gives rise to certain meanings and practices, 2) the unintended consequences of actions, 3) structural conflicts within organizations and societies, and 4) the temporal dimension of a social order and how any text must be seen in its historical context (Myers 1995).

The hermeneutic cycle summarizes the basic analytic process (See Figure 3). In a study of trace data, the researcher would constantly move back and forth between the whole corpus and its parts. This cyclical movement allows the researcher to start with a vague understanding of the whole text only to interpret its different parts. Out of these partial interpretations the parts are related back to the whole, and so on. This cyclical movement implies a continuously deepening understanding of the traces by contextualizing once current interpretation of the

whole with further detailed analysis of selective parts. New insights are then integrated as the researcher expand the perspective towards to global meaning. The analytic process implies a constant comparison between interpretations of small segments and the global meaning.

Figure 3. Hermeneutic Cycle



The need for an interpretive approach is clearest when dealing textual traces. For example, we might be interested in how volunteers use talk to support discovery, such as the discovery of the convict worm in Seafloor Explorer. Simply counting posts is unlikely to be satisfactory. Instead, the researchers would read and reread messages to form an interpretation of the kinds of messages and their function and then test that growing understanding against a larger set of messages. For example, we could examine how a volunteer calls attention to an apparent discovery, how other volunteers respond, the kind of evidence that attracts the interest of the science team and so on, to build a theory of volunteer-led discovery. In this case, the hermeneutic approach is applied much as in any qualitative study.

While the need for interpretation is clear for qualitative data, we note that an interpretivist approach can also be useful also for forming an understanding of the meaning of quantitative trace data taken from an online system. The data record the various actions that users take on the site, but to understand the import of these data requires understanding the purpose and meaning of the captured interactions.

Applying a hermeneutic approach to trace data faces complications. As noted above a typical first step in the hermeneutic cycle is to pick a subset of data on which to focus. However, picking which trace data to read needs to be done carefully. For example, if participation is skewed, as is often the case in online communities' data, a random sample of users will likely include many users who are not very active and may miss particularly active users. An alternative is a stratified sample: e.g., to group events by user and to select representative users at different levels of activity or even to focus on the most active.

The next step in the cycle is to develop an interpretation of the sample of data. At the most basic level, the researcher needs to understand the mapping of actions that a user can take on the system to the data that are recorded in the traces. While data may have labels (e.g., in a database dump), the connection between that label and an action is not always straightforward. For example, a database architect might give a database field a name that is suited to the context of the system's creation, but which may not be meaningful to a user or a researcher looking from the outside. Such an understanding can be built initially by engaging in the system personally, as in virtual ethnography, and the understanding refined through the hermeneutic cycle.

Technologies are often used differently than intended by the designers, so it is important to understand how users enact the system in practice, and what the recorded system actions mean to users. For example, what exactly does a user mean when they click "Like" in Facebook? The actual meaning of those events and associated data can differ from a common-language understanding of the label. Geiger and Ribes (2011) call the process of learning the meaning of digital traces "inversion". Complicating things further, different users may mean different things by their use or use a feature with different levels of intensity. And yet, to assign meaning to the trace data, these nuances must be understood.

A key point of a hermeneutic approach is that to decode the meaning of a trace, it must be understood within the broader context of the platform that captures the activity. However,

trace data often lack situational clues, so it takes work to establish the context of the events. It may be useful to compare across time, settings or projects or to position traces in context with other work, perhaps other activities happening at the same time. From the small-scale interpretation, the researcher gradually builds up a fuller interpretation of system usage as a whole. As with any hermeneutics analysis, different traces may provide stronger or weaker evidence of theoretical constructs of interest. The evidence may sometimes be direct: e.g., friending someone on social media site suggests a certain attitude towards that person. Often though, the evidence may require more inference.

3.3 Diffractive Approach: Building an apparatus

We turn next to *diffractive* approaches to analyzing trace data. From a socio-material standpoint, even a hermeneutic approach falls short on at least two counts. First, the analysis assumes the existence of a pre-given whole, a corpus of trace data to be analyzed; the hermeneutic steps do not question the completeness of this body of data. Second, they also do not consider the potential effects the analytic process might have on traces, the information system, and the users. More recent pragmatic and critical hermeneutic approaches do consider the researcher's position vis-a-vis the study subject and historical changes an information system might go through. Yet, even with this consideration, traces are still treated as reflections of users' practices. The diffractive processes involved in the making of these traces remain shrouded in a reflective methodology.

Probing deeper into the socio-material nature of online systems, we recognize that the trace data we seek to analyze are collected by a system and thus shaped by its design. From this perspective, trace data are not pre-given entities, packaged and waiting to be picked off a shelf. Rather, traces are the product of (and so entwined with) an apparatus, a material configuration of the word, which iteratively reconfigures traces, information systems and users as part of the ongoing process of becoming (Barad 2007).

This perspective leads us to recognize that before researchers can start analyzing trace data, any data, they need to start building or extending an apparatus. In the case of our study of the Zooniverse, an apparatus was already partly configured when we started our work, in the form of the Zooniverse platform described above. Yet, for the researcher, building of the apparatus does not end with the traces captured by an existing information system.

First, trace data needs to be collected and further reconfigured. Data might be generated by "scraping", that is, recording data from publicly visible web pages. With cooperation of system managers, it may be possible to obtain dumps from the databases driving the system. In either case, capturing and managing a volume of structured data often requires a different kind of data analysis infrastructure than is needed for analyzing typical research data or textual documents. For example, we found that needed their own versions of the various database systems to process the dumps and that data analysis required considerable preprocessing to get the data into a usable format.

Second, while it is tempting to expect that the system captures traces of all events, the perspective of system as apparatus reminds us that data storage is itself a practice, and the assumption of completeness must be carefully examined. Despite the hard work and best intentions of system developers, one cannot assume that the apparatus captures all relevant activities. For example, as Howison et al. (2011) point out, systems are subject to many problems that result in data loss (e.g., server outages, disk failures, deleted log files, or truncated database tables), meaning that trace data—even from database dumps—may be incomplete, though the problems may not be immediately visible. To address these problems requires developing a deep understanding of the fine details of the technical system. Unfortunately, it can be hard to obtain the necessary exposure without the assistance of those running the system (Howison et al., 2011).

In other cases, activities of interest may be unavailable for integration into the researcher's apparatus. For example, the user interface available for scraping may show only a

subset of data (e.g., omitting some historical or private data or displaying only a sample of a large dataset). If a citizen science system is designed primarily to support the science, it might only record the annotations done and not activities such as studying the tutorial that the designers do not consider to be data (as in Zooniverse when we first started our studies). Likewise, many systems only record activities after a user logs into the system. This limitation excludes from trace data the work of volunteers who do not bother to log in.

Important activities might also take place that are not performed with the system and so are impossible to capture in any system trace. For example, in studying citizen science, trace data will capture annotations and some of the discussion, but not the individual work done by volunteers to analyze an image or resources consulted outside the system. We found cases where citizen science volunteers undertake work using entirely different systems. For example, in some of the projects we studied, volunteers carry out follow-up analyses of scientifically-interesting objects using Google Spreadsheets and discuss them in periodic Google Hangout meetings. None of this analysis is captured in the Zooniverse system. The same caveat applies to other kinds of online contribution, e.g., work done for debugging or system design in open source development or research done to support Wikipedia editing. These omissions imply that studies using trace data need to consider critically the boundaries of the apparatus and the activities it configures.

Third, describing complex events might require data from multiple sources. For example, making sense of a citizen science annotation required merging together data from multiple sources, such as records of user data, metadata, and the object, and these may be stored in different databases and database tables. Such merging poses its own problems. For example, it is common to want to assign traces to individual volunteers. However, a single individual may have multiple representations in the traces (e.g., variations in spelling of a name or multiple login IDs in different systems) or conversely, multiple individuals may share a single ID. On

some systems, a subset of actions can be performed by users without logging in, so the actions associated with a user ID may not include every action performed by that individual.

3.4 Diffractive Approach: Process

A diffractive analysis approach further invites us to consider the temporality of the data. The apparatus built to capture traces does more than projecting distinct units, whether single actions, individuals (all activities performed by one person), groups, project or the whole system. A diffractive reading allows us to explore processes, the iterative configurations and reconfigurations of socio-material practices involving the users, the system and the traces. The focus is not just on "what is" but on the ongoing dynamics of becoming. Temporal patterns emerge out of a diffractive reading.

For example, in many of our studies, we seek to understand user's practices by examining the work done in sessions. The intuition is that users will often interact with an online system for some period, creating a temporally-adjacent set of traces, then take a break (e.g., until the next day). Traces of events separated by a short gap can be grouped together in a single session, separated from the next session by a longer gap. This analysis approach provides a unit of analysis between an individual trace data item and an entire life history that acknowledges the temporality of a volunteer's interactions with the system.

Creating sessions from raw trace data require refinement of the apparatus. First, we have to decide how to choose the set of activities that comprise a session from the stream of activities recorded in the system. Prior work on Wikipedia had defined a gap of one hour between activities as indicating the start of a new editing session, but given our understanding of the nature of the work in Zooniverse, built from our own interaction with the system, and through observation of the distribution of gap sizes, we chose a gap of 30 minutes instead (that is, the sequence of activities separated by less than 30 minutes were considered a session).

Second, creating sessions required aggregating data from multiple events in the trace data. An important consideration in aggregating data is picking which data to keep to represent the aggregation, and which details to suppress. For example, a session might be represented by counts of different kinds of actions, but doing so loses information about the ordering of events. Other researchers apply sequence analysis techniques to analyze this order information (e.g., Keegan et al., 2015).

Third, sessions with similar patterns of activities can be identified statistically using cluster analysis. However, decoding the nature of the sessions in those clusters required a diffractive reading of quantitative analysis. In this case, we tried to understand what the sequence of activities looks like as carried out by a user. The different clusters of sessions exhibited distinctive patterns of work, e.g., sessions focused primarily on annotating vs. sessions that included a mix of annotation and talking. Different kinds of users were then identified by the mix of session types they exhibited throughout their span of interaction with the system. The researchers thus moved through analysis at the individual activity level, to the session level, and then the user history level in their analysis.

The final analysis required a back and forth comparison between, on the one hand, individual activities and streams of activities within a session, and on the other, patterns exhibited in the different clusters. By exploring the interferences among the many traces, we found, for instance that while only a handful of users regularly contribute to talk, many others refer to those discussions in their sessions and seem to find them useful in structuring their unfolding project work.

3.5 Diffractive Approach: Reading insights through one another

Finally, a diffractive approach opens the possibility to read elements of the apparatus *through* one another. The traces are not merely reflections of user practices: the traces diffract back to the users, the information system and the researcher. The apparatus serves not only

research purposes: rather, it is a medium through which actions and interactions takes place and so shapes what is possible. The apparatus plays a constitutive role in the production of the phenomenon by enacting specific boundaries in the socio-material reality. To extend Barad's metaphor, it is as if the experimental apparatus reached out and changed the light source to behave in some ways and not in others (rather than just changing the light passing through). For example, Zooniverse projects never allow volunteers to see other volunteers' annotations and provides access to talk about an object only after the user submits an annotation, both to avoid propagation of user biases. These constraints mean, for example, that legitimate peripheral participation is problematic as a source of learning for newcomers. Volunteers have no easy way to observe more advanced volunteers' work practices (though the talk posts noted above provide some partial compensation).

One can explore *through* different aspects of the system by engaging with the apparatus in different ways. First, the researchers can change the system and so affect user behaviour even as they are using the system to study those behaviours. Various quasi-experimental approaches involve tweaking how the system presents objects to volunteers and so the way the volunteers interact with the objects. The researchers then seek to understand how these new traces structures diffract back onto volunteers' practices. The large volume of available data would enable researchers to see effects despite the large variability in volunteer behaviours.

In a new Zooniverse project, for instance, we implemented multiple levels of training (augmenting the usual tutorial and field guide). In this system, volunteers annotate images for the class of object shown in the image (as in Snapshot Serengeti). However, rather than providing all classification options to new users, the system introduces them a few at a time, using a machine learning classifier to identify images that were likely to be of those classes and so good exemplars to learn from. The system is thus shaping the motivation and ability of the volunteers through their interaction. In other experiments, we manipulated the presentation of

images to appeal more strongly to different motivations, again shaping the behaviours as well as recording them.

Second, direct engagement with volunteers offers ways to explore the apparatus and its diffractive patterns. Participant observations and interviews with individuals and in focus groups allow researchers to compare personal experiences with trace data. Trace data can also be helpful to ground interview questions in actual recorded behaviour. This process involves more than the simple triangulation of one statement against another statement. By using traces to structure participant observation and interviews, the researchers can explore the inner workings of the apparatus. In the process, they may learn as much about the volunteer's practices as about the way traces are recorded or not.

Thirds, one cannot forget that we as researchers are an integral part of the apparatus; not in the sense that we distort some reflection of user behaviours, but rather, that our active engagement in the building and running of the apparatus offers rich opportunities to explore its inner workings and what it allows us to know and not to know about the ongoing dynamics of becoming associated with the system. Data collection practices are open to rearrangements and the creativity of scientific practices includes the skill of making an apparatus work for a purpose. Elements are reworked and adjusted, leading to adjustments of the boundaries and cuts performed by the apparatus and the nature of the phenomenon.

In some cases, pre-existing traces might be sufficient to address the research phenomenon. But when traces are incomplete or fail to completely address the behaviours of interest, additional data are needed. One possibility is creating additional traces. The researcher might play a role as a co-constructor of traces, arranging with software developers to have the system collect new traces. For example, through our interaction with the Zooniverse developers, the data collected by the system has changed, in part in response to our interest in analyzing more aspects of user behaviour. The expansion can be iterative where the researcher cycles

between appreciation of what new data can be collected (or is able to be observed) and consideration of appropriate sources of evidence to address the phenomenon.

In our own work, we have spent considerable time defining specific data related to users' interactions online. As an example, lurking (using a system without visible participation) is a common step in a user's learning how to participate in an online community. However, lurking is often not observable in trace data. System developers had not considered the possibility of analyzing anonymous users, so the ability to track their behaviour did not exist. Adding a capability to track which webpages users visited created novel trace data that was constituted for a study of learning in Zooniverse. Thus, researchers can play a central role in building the apparatus and its resulting traces that can go beyond cobbling together parts from existing systems.

Many other researchers will not have the same opportunity to influence system design. Nevertheless, even a study of a fixed systems, e.g., analyzing Twitter data, still involves the building of an apparatus. For example, a researcher's apparatus might be narrowed to only include tweets for a single event or a specific period. Most of the parts of the apparatus are designed by others but the researchers make informed decisions and the process involves the creative skill of making the apparatus work for their purpose.

Finally, despite our hand in the building of the apparatus, it is important to realize that we do not have a god's eye view that allow us to see the whole system and its interactions with users and the environment (diffractive or not). In studies of online systems in particular, most of the users and their interactions are not accessible to us, because they are interacting from their own settings. Taking Barad's metaphor of the two-slit experiment, we cannot see the light source and the slits and trace the path of the light in an apparatus laid out on a lab bench. Only by reading insights through one another can we piece together a better understanding of the iterative reconfigurations that are part of citizen science practices.

4. Principles: A diffractive reading of trace data

Facing a continuous torrent of trace data, IS researchers confront a number of methodological challenges. Table 1 summarizes seven guiding principles for a diffractive readings of trace data. These are not bureaucratic procedures to be followed one after the other but fundamental ideas that guide the research process. We argue that researchers will be better served by avoiding the shortcomings of approaching trace data as mere reflections of a pre-given reality. A diffractive reading offers a process perspective to trace data, highlighting the ongoing and socio-material dynamics of becoming. Trace data stand out not merely as a data set or text to be collected and dissected but the product of an apparatus cobbled together by the researcher with help from system designers and users in an effort to demarcate a phenomenon of interest. Likewise, the interpretive work associated with trace data takes on a performative character. Building the apparatus associated with trace data analysis highlights the nexus of doings, materialities, and discourses that researchers enact to support their investigation. There is a design element to trace analysis (Bjørn & Østerlund 2014). The boundaries defining the phenomena of interest are not pre-given subject and object. Instead the researcher needs to pay careful attention to how the building of the apparatus demarcates different entities and the way they co-constitute one another.

Table 1. Principles Guiding a Diffractive Analysis of Trace Data
1. Diffraction not reflection Trace data are not mere reflections of some external reality made up of pre-given substances. Instead trace data are part of diffractive processes of becoming. Multiple processes interfere and carry with them forward the history of prior encounters.
2. Sociomaterial The social and the material are inseparable. What is social is also material and vice versa. Sociomaterial formations should be approaches as co-constituted through a nexus of doings, materialities and discourses.
3. Building an apparatus The research process involves the building of an apparatus. This building is an ongoing endeavor that stretches across the entire study. The apparatus includes not only an information system but also trace data, users, designers and researchers, all of which are co-constituted.

4. Boundary making The phenomena of interest are not pre-given subject or objects but co-configured through the research process. Boundaries are made through the building of the apparatus, which allows the researcher to demarcate the phenomena of interests, participants, trace data, information systems, etc.
5. Temporality With an emphasis on processes, temporality plays a key role in our understanding of 1) the research process and 2) the ongoing dynamics through which a phenomenon becomes.
6. Reading insights through one another Insights are read through one another. This reading through allows the researcher to explore interferences in the apparatus and differences that matter. These readings can be both empirical and theoretical in nature. Empirically, trace data through their information systems. Theoretically, quantitate analysis or hermeneutic process can be re-configured through a diffractive reading.
7. Cyclical Analysis Diffractive analyses are iterative or circular in nature. Insights about one element of the apparatus are contextualized in regard to insights from other parts of the apparatus. This allows for an iterative integration and expansion of insights about not only the workings of the apparatus but also its boundaries

Temporality is a particularly important concern in trace analysis, both for understanding the analytic process and for our conceptualization of the phenomena at hand. First, diffractive readings emerge out of recurrent and cyclical analysis. As has long been recognized by hermeneutic methodologies, it takes repeated readings to deepen understanding. for a diffractive methodology, this process also involves the gradual configuration and reconfiguration of an apparatus. Second, trace data help track the ongoing dynamics through which a phenomenon becomes, such as through the evolving activities and infrastructure on a citizen science site like Zooniverse. The temporal order of social phenomena is well documented in the socio-material literature, which often focuses on the continuous entanglement of the human and the material over time (Cecez-Kecmanovic et al., 2014). Equally important, the apparatus often changes over time in ways that are not always under the control of the researcher. These changes force researchers to pay careful attention to the provenance of the traces and the history of the apparatus they build.

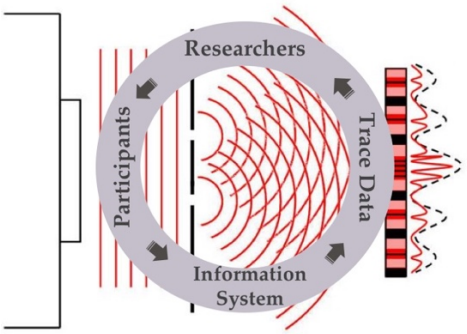
A diffractive methodology does not leave quantitative and qualitative techniques to the dustbin, however. As we notice in our diffractive approach to citizen science practices in Zooniverse, techniques such as cluster analysis and hermeneutics still have a role to pay. But

one cannot rely on them to simply reflect pre-given objects. Instead, reading them through a diffractive approach offers new insights to the building of an apparatus and the understanding of online phenomena. As illustrated in Figure 4, a diffractive methodology integrates insights from the hermeneutic process, in which the researcher constantly contextualizes, integrates and expands partial analysis by reading them through other parts of the research phenomenon. However, in a diffractive approach, the analysis does not move between parts and an assumed whole. Instead it reads elements of the apparatus through one another and the phenomena of interest. This approach involves boundary making, or in Barad's (2007) terms, agential cuts.

Studying Zooniverse, one cannot assume that traces easily scrapped from the system constitute a whole. The analyst needs to read the traces through the socio-material arrangements of the information systems supporting Zooniverse. The trace data should also be read through the practices of the volunteers and the researchers. Other systems, practices and traces may emerge as central to the phenomena. Likewise, the volunteers' practices can be read through the traces, information system and activities of the researcher. Volunteers are not simply leaving traces behind them as a boat cutting a wake in its path. The traces make up part of the reality that define the volunteers' activities. What one sees on the Zooniverse site is partly a product of one's own and other volunteer's traces. The boat is rocked by its own wake as it plows through a canal, with each wave diffracting back to the boat after hitting the channel banks. The diffractive patterns of the waves must be read through the rocking of the boat, the structure of the embankments and the decisions of the pilot trying to avoid spilling his morning coffee. The diffraction pattern mark differences that matter.

The analytic cycle depicted in Figure 4 should be taken as a set sequence suggesting that the traces should be read through the information system, which should be read through the volunteers, and so on. The figure suggests how a researcher should read the different elements through one another in an ongoing analytic process where the researcher contextualizes, integrates and expands insights on the ongoing becoming.

Figure 4. Cyclical Analysis



Cycling through the apparatus as depicted in Figure 4 holds parallels to Nicolini's (2013) notion of zooming in and out of practice. Drawing on a botanical metaphor and work by Deleuze, Nicolini compares the research process to the way some plants spread rhizomes, an expanding, nonhierarchical web of roots. In this perspective, a study starts by inquiring about one location (zooming in) and then extends the analysis to other locations following emergent relations (zooming out). Continuing this iterative process enables scholars to follow lines of becoming and to describe how boundaries are formed and taken apart. By observing and experiencing the phenomena from different angles, the dynamics of practice exposes the becoming of technologies, people and entities, and how their boundaries and properties are reshaped, with what consequences and for whom (Cecez-Kecmanovic et al 2014: 821).

Zooming in is a theoretical act, and different practice theories define practice differently. Thus, Nicolini suggests that each re-zooming involved the application of a different practice theory. A diffractive reading offers a number of additional insights, which adds to our socio-material exploration of trace data. First, a diffractive reading take as its starting point the apparatus and the way it iteratively reconfigures space-time matter as part of the ongoing dynamics of becoming. In trace analysis, we cannot ignore the socio-material matter involved in ongoing making of these traces. Second, dealing with an apparatus require a careful reading of each element through other parts. This approach to analysis allows the researcher to look for differences that matter. That is, practices send ripples through the apparatus and in the process,

build synergies, cancels each other out or produce new ways of being. Diffractive waves carry with them the history of the source and the apparatus.

Diffractive analysis offers a renewed take on the notion of engaged scholarship, originally defined by Van de Ven's (2007) as research striving to obtaining different perspectives of key stakeholder in the study of complex problems (e.g., users, developers or researchers). These interactional views typically strive to bridge professional and research practice. A diffractive analysis engages not only users, developers and the researchers but equally important, traces and information systems in a cyclical exploration. This engagement tinkers with their mutual boundaries and temporality. Attention to the apparatus of research thus helps researchers achieve the promise of trace data for a richer understanding human behaviour.

5. References

Barad, K. (2003). Posthumanist performativity: Toward an understanding of how matter comes to matter. Signs, 28(3), 801-831.

Barad, K. (2007). Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning. Duke University Press.

Bjørn, P., & Østerlund, C. (2014). Socio-material -Design: Bounding Technologies in Practice. Switzerland: Springer.

Butler, T. (1998). Towards a hermeneutic method for interpretive research in information systems. Journal of Information Technology, 13, 285-300.

Cecez-Kecmanovic, D., Galliers, R. D., Henfridsson, O., Newell, S., & Vidgen, R. (2014). The socio-material ity of information systems: current status, future directions. MIS Quarterly, 38(3), 809-830.

Cole, M., & Avison, D. (2007). The potential of hermeneutics in information systems research. European Journal of Information Systems, 16(6), 820-833.

Gadamer, H. G. (1975). Truth and Method, trans. W. Glen-Dopel, London: Sheed and Ward.

Gaskin, J., Berente, N., Lyytinen, K., & Yoo, Y. (2014). Toward Generalizable Socio-material Inquiry: A Computational Approach for Zooming In and Out of Socio-material Routines. MIS Quarterly, 38(3), 849-871.

Geiger, R. S., & Ribes, D. (2011). Trace ethnography: Following coordination through documentary practices. In Proceedings of the Hawaii International Conference on System Sciences (HICSS). doi: 10.1109/HICSS.2011.455

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. Journal of the Association for Information Systems, 12(12), 767. Available from: http://search.proquest.com/docview/916253418?accountid=14214

Jackson, C., Østerlund, C., Maidel, V., Crowston, K. & Mugar, G. (2016). Which Way Did They Go? Newcomer movement through the Zooniverse. In Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16).

Keegan, B. C., Lev, S., & Arazy, O. (2015). Analyzing Organizational Routines in Online Knowledge Collaborations: A Case for Sequence Analysis in CSCW. In Proceedings of the 2016 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '16). ACM, New York, NY, USA, 257–266.

Kvale, S., & Brinkmann, S. (2009). Interviews: Learning the craft of qualitative research interviewing. London: Sage.

Leonardi, P., Nardi, B., Kallinikos, J. (2012). Materiality and Organizing: Social Interaction in a Technological World. Oxford University Press, Oxford.

Lyytinen, K.J. and Klein, H. (1985) The critical theory of Jurgen Habermas as a basis for a theory of information systems. In Research Methods in Information Systems, Mumford, E., Hirschheim, R., Fitzgerald, G. and Wood-Harper, T. (eds) Amsterdam:Elsevier Science Publishers

Myers, M. D. (1995). Dialectical hermeneutics: a theoretical framework for the implementation of information systems. Information systems journal, 5(1), 51-70.

Nicolini, D. (2012). Practice theory, work, and organization: An introduction. New York: Oxford University Press.

Orlikowski, W., Scott, S. (2008). Socio-materiality: challenging the separation of technology, work, and organization. Acad. Manage. Ann. 2(1), 433–474.

Ricoeur, P., & Thompson, J. B. (1981). Hermeneutics and the human sciences: Essays on language, action and interpretation. Cambridge university press.