Governing *PatientsLikeMe*: information production and research through an open, distributed and data-based social media network

Forthcoming on *The Information Society*; accepted July 15th, 2014.

Niccolò Tempini, University of Exeter

Running title: Governing *PatientsLikeMe*

Contact Information: Niccolò Tempini, EGENIS, Department of Sociology, Philosophy and Anthropology, University of Exeter, Byrne House, Exeter, United Kingdom EX1 4PJ, EMAIL: n.tempini@exeter.ac.uk, URL: www.tempini.info

Abstract

Many organizations develop social media networks with the aim of engaging a wide range of social groups in the production of information that fuels their processes. This effort appears to crucially depend on complex data structures that afford the organization to connect and collect data from myriad local contexts and actors. One such organization, *PatientsLikeMe* is developing a platform with the aim of connecting patients with one another while collecting self-reported medical data, which it uses for scientific and commercial medical research. Here the question of how technology and the underlying data structures shape the kind of information and medical evidence that can be produced through social media-based arrangements comes powerfully to the fore. In this observational case study I introduce the concepts of *information cultivation* and *social denomination* to explicate how the development of such a data collection architecture requires a continuous exercise of balancing between the conflicting demands of patient engagement, necessary for collecting data in scale, and data semantic context, necessary for effective capture of health phenomena in informative and specific data. The study extends the understanding of the context-embeddedness of information phenomena and discusses some of the social consequences of social media models for knowledge making. .

Keywords: self-reporting, social media, data-based, distributed science, aggregation, information cultivation, social denomination, semantic context, engagement, PatientsLikeMe

Introduction

Organizations developing social networking sites (boyd and Ellison, 2008), by offering new kinds of information services to a user base of unprecedented scale, can explore new data-based (Aaltonen and Tempini, 2014) business models centered on the collection, analysis and repackaging of data generated through network infrastructures (Aaltonen and Tempini, 2014; Kallinikos, 2006; van Dijck, 2013; boyd and Crawford, 2012; boyd and Ellison, 2008; Kallinikos and Tempini, 2011; Mayer-Schönberger and Cukier, 2013). Typically these systems routinely produce information from the data that users generate while dealing with the matters of their own lives. As noted by Howe (2008), the capillary reach of these networks captures the ephemeral but valuable knowledge of diverse and distributed local contexts, which tends to escape universal models (Hayek, 1945). Nonetheless, use of information technology to connect to diverse local contexts that were previously out of reach reconfigures, rather than solves, the tension between the universal, standard models and the specific contextual instances they ought to relate with (Agre, 1992; Berg and Timmermans, 2000; Bowker and Star, 1999). In this respect, the reliance of social media technologies on complex data structures reproduces the reductive operational logic of selection, identification and classification. As we enter an age of intermediated, data-based and standardized community life (Bowker, 2013; Kallinikos and Tempini, 2011), understanding the mechanisms that shape the development of social media and the data structures that power them is of paramount importance.

In this paper, I analyze the case of the Cambridge, Massachusetts based organization *PatientsLikeMe*. The for-profit company, founded in 2004, has been developing an ad-free social networking site whereby patients can connect with each other as they collect self-reported medical data.[1] The research team exploits the collected data for scientific and commercial medical research purposes. To date, the research on *PatientsLikeMe* includes 37 scientific publications, based on data contributed by more than 220,000 patients. Contributions in peer-reviewed articles, conference papers, reports, and editorials have covered many different facets of *PatientsLikeMe*. To give just a few examples, an article published in *Nature Biotechnology* (Wicks *et al.*, 2011) disproved through a virtual clinical trial the efficacy of lithium carbonate for Amyotrophic Lateral Sclerosis (ALS) patients. Another article (Wicks and MacPhee, 2009) assessed the prevalence of social issues (compulsive gambling) in the Parkinson's disease patient population by comparing it to another patient population dealing with ALS, a chronic progressive neurological disorder, in order to test hypotheses on the emergence of this association – a difficult comparison to achieve. Other works have looked at symptom distribution discoveries (Turner *et al.*, 2011; Wicks, 2007) and the relationship between patients' and experts' language regarding health experiences (Arnott-Smith and Wicks, 2008).

PatientsLikeMe styles itself as providing an all-encompassing platform for the organization of patient sociality and advocacy. It aims to become the social media network of choice where relationships between patients, clinical professionals, healthcare providers, pharmaceutical companies, patient organizations and NGOs (non-governmental organizations) are discussed or intermediated. In this sense, *PatientsLikeMe* differs from patient and evidence-based activism organizations (Epstein, 2008; Rabeharisoa *et al.,* 2013). It is a new kind of intermediary. Critically

[1] More information can be found at www.patientslikeme.com/about/

depending on patient involvement and observation and research skills, it is a champion of the most recent participatory turn in medicine (Prainsack, 2014). At the same time, because of how the data are controlled and the way the organization's business model is designed, most of the research the network has produced has been dependent on the occasion of related commercial research projects.

For understanding an innovative organizational form such as that represented by *PatientsLikeMe*, it is critical to explain the conditions that shape the production of information out of data. In the case of *PatientsLikeMe,* researchers do not learn about the patients, their experiences and their health situations in any other way than through the social data. The social media infrastructure of *PatientsLikeMe* is therefore the cognitive grid through which the world is captured, represented and read (Kallinikos, 1999; Ribes and Bowker, 2009; Bowker and Star, 1999; Zuboff, 1988).

Research has focused on how social media afford new organization forms for knowledge production (Treem and Leonardi, 2012), facilitate exchanges within or beyond organizational boundaries (Majchrzak *et al.*, 2013), and support the generative liveliness of seemingly self-organized online communities (Faraj *et al.*, 2011). In general studies have emphasized how these networks link users, content, and combinations of the two (Treem and Leonardi, 2012), but have not unpacked the role data structures and models play in the construction of these connections. For our present analytical project, it is critical to understand the often-invisible work processes and devices that make data comparable and translatable across contexts (Star and Lampland, 2009; Star, 1983, 1986).

Technical structures (data, protocols, algorithms, software) shape our understanding of both local and distant contexts through selective and ordered representations of the world (Berg and Timmermans, 2000; Bowker, 2013; Williams, 2013), making it possible to count and describe distributed phenomena – operationalizing new sets of unifying and dividing practices (Bowker and Star, 1999; Rose, 1999, 2007). To represent knowledge in data structures means to articulate in practice what Leonelli, in the case of bio-ontologies, calls 'classificatory theories' (Leonelli, 2012). According to Leonelli (2012, p. 58), information infrastructures for scientific collaboration embed theories as they 'aim to represent the body of knowledge available in a given field so as to enable the dissemination and retrieval of research materials within it; are subject to systematic scrutiny and interpretation on the basis of empirical evidence; affect the ways in which research in that field is discussed and conducted in the long term; and—most importantly if we are to regard them as theories—express the conceptual significance of the results gathered through empirical research.'

Issues of ontological representation are not simply a theoretical dispute. They are in fact grounds for political struggles of representation of social objects and subjects. The outreach and involvement of the target community is essential for achieving the cross-contextual adoption and knowledge integration for which an information infrastructure is built. To be successfully adopted, a system developed for a distributed patient user base must be recognized as faithfully representing the knowledge of the community of reference (Millerand and Bowker, 2009; Ribes and Bowker, 2009; Ribes and Jackson, 2013). This can be particularly difficult to achieve in social media networks, where the user base is at the same time open, undefined, and of inherently uncertain availability. Moreover, the data structures in *PatientsLikeMe* are subject to systematic scrutiny only between the organization and the research partners, as their limited visibility from outside – embedded in the workings of the system – does not facilitate further warrant. Thus knowledge representation and embedment in information infrastructures is matter of political struggle especially in contested or evolving knowledge domains. It is not 'simply a matter of properly capturing knowledge but also a question of whose knowledge to capture' that is at stake (Ribes and Bowker, 2009:210).

We need to also consider the techniques of collection and analysis themselves. Building on Bateson's definition of information as 'difference that makes a difference' (Bateson, 1972), Jacob (2004) compares between systems of categorization and classification, distinguishing by the different degrees of semantic context and flexibility to local context they express. By semantic context Jacob refers to the information that is embedded in the structure of a data model, and expressed by the degrees of differentiation between semantic fields that the structure expresses with the shape of its own organization. A more structured data model embeds more information, because its ability to differentiate between phenomena and relate them to other data is greater (Bateson, 1972; Jacob, 2004; Kallinikos, 2013). However, more structured systems (with richer semantic context) are less flexible in terms of being used for specific local contexts. Conversely, systems that are less structured are more easily adapted to local practices and situations.

Against this backdrop, *PatientsLikeMe* with its massive involvement of an open and distributed user base via social media offers a good site for the exploring following questions: (1) How are the data structures developed to carry reliable information out from the patient life context and to the researchers' in a way that satisfies the requirements for medical scientific research? (2) What factors shape the amount of information that can be expressed by data collected through an open, distributed network? (3) How is the patient user base governed to select and encourage desired behavior? With this exploration, the intent is to deepen our understanding of 'semantic gateway technologies' (Ribes and Bowker, 2009), which translate knowledge between the organization and a myriad of local contexts.

This paper is structured as follows. In the following section, I briefly describe the methodology of the case study, explaining how I selected and worked through the empirical evidence. Next, I present the empirical evidence, by providing first a short overview of the organization, then an analysis of a short series of observed, topical events of information cultivation that emerged from the case as requiring a theoretical explanation. Finally, I discuss the evidence, elaborating a theory of information cultivation in open and distributed networks and pointing out major implications for the understanding of social media organizations and Internet medical research.

Methodology and research design

For 26 weeks – from September 2011 to April 2012 – I conducted an observational case study at the headquarters of *PatientsLikeMe* in Cambridge, Massachusetts. I worked as a member of the R&D and Health Data Integrity teams and participated in work activities, through regular working hours, five days a week. I was fully involved in projects, also occasionally represented the organization at conferences, meetings and conference calls.

Data collection included a number of different sources of data, enabling robust triangulation for construct validation (Yin, 2009). In addition to interviews, and the observation of meetings and work processes, I was allowed to access work documents in various formats, and to take screenshots on both the admin and the user side of the system. With no monetary exchange being involved, I was free to considerably modulate my effort and participation. My role allowed me to exercise a great degree of discretion over my commitments. I had more freedom than regular employees to regulate my involvement in projects. I could take frequent breaks, when I needed to make notes. I had extensive access to organizational resources, and I was able to obtain more resources when needed. The flexible nature of my participation in the organization enabled me to work with most of the employees based at the company's headquarters – about 30-40 people, including turnover. I participated in numerous meetings, including one-to-one meetings, project-specific team meetings, regular weekly team meetings, company meetings, 'stand-up' agile development meetings, and release demo meetings.

I interviewed the great majority of the employees of the company, at all levels of the hierarchy. I concentrated most of the interviews towards the end of my fieldwork period, interviewing some participants a second time if necessary. In this way, I was able to focus the interviews on specific topics, based on the observations collected to that point, and to test more developed hypotheses. Interviews were a primary means for validation of emerging explanations (Runde, 1998). Running the bulk of the interviews at the end of my fieldwork period allowed me to have clearer knowledge of my interviewees' work roles and expertise. An interview guide was developed anew for each of the semi-structured interviews.

During the fieldwork period, I developed tentative interpretations of the phenomena I had been observing. In my time off-site (evenings, weekends), I reviewed and further integrated my notes (Mingers, 2004; Sayer, 2000). I used retroductive reasoning, wherein starting from the observation of an event that requires an explanation a hypothetical cause is fitted *post hoc* to fill the knowledge gap (Mingers, 2004). Hypothesized causes do not need to wholly account for the observed event, and they can also have a varying ability to repeat their effects in an observable fashion, as countervailing powers might oppose their manifestation (Runde, 1998). Here relevance is more important than regularity (Runde, 1998; Sayer, 2000). The events that attracted my attention could be small and ephemeral, such as fleeting comments, or big and noticeable, such as unexpected systems development decisions (Wynn and Williams, 2012).

I logged all these reflections in a separate electronic log and I used tags as provisional codes, to aid my recollection of events and topics. Also, I used my time away from the office to research literature that could help me formulate hypotheses about the phenomena I was witnessing. I kept the logs, with narrations of events as I experienced them as well as interpretations, accessible to me at all times during the fieldwork. When preparing for each interview, I scanned through these logs and reviewed the points I was developing to aid my discussions of phenomena of interest with the interviewees. After the fieldwork the analysis stage, I started to converge all the pieces of evidence to compose the analytical narrative that I share in this paper. Analytical writing, in its various stages, is not only a process for grounding an argument that needs to be demonstrated. It is itself a technique for facilitating retroductive theorizing (Aaltonen and Tempini, 2014).

As an initial approach to conducting the research, I began the fieldwork with the aim of understanding the role of technological structures within the organizational setting, with particular regard to the forms of knowledge representations embedded in data structures and how such structures shape the data collection tasks and the real-world medical evidence that the organization is able to produce. As I argued in the introduction, this research combined an exploratory research question with an innovative empirical setting. Intensive observational case studies are a well-suited methodology for this kind of research design (Yin, 2009). They allow to build new theory while taking into consideration the whole complex of factors that make up an empirical setting (Sayer, 2000).

The tension between patient engagement and semantic context, data scale and specificity, emerged in the field as a recurrent issue in the management and development of the system. Soon, I started to formulate provisional interpretations of the observed phenomena and I searched the literature for frameworks that could guide my observations. Initially, I was inspired to interpret the tension in terms of the continually moving boundary between the aspects of the world that are modeled in a technology's constructs and rule-bound behavior (the *'order'*), and the opposing *'disorder'*, namely the aspects of the world that technological constructs ignore, as proposed by Berg and Timmermans (2000). They argue that a technological *order* can sometimes be more successful in achieving universal application when it stipulates behavior or models the world less, instead of more, in its constructs. A compelling and instructive argument, it soon became clear to me that this one-dimensional characterization was too abstract for the empirical setting of this research. Understanding the development of a complex system such as *PatientsLikeMe* in terms of the shifting boundary between the fields of order and disorder was not helping me to explain the specific drivers and effects of change. The risk was that I might analytically blackbox the technology and fail to look into its components and their interrelationships. I started formulating endogenous explanations, closer to the empirical reality I was observing, guided by the critical realist framework. This was also necessary as it created a common ground for my conversations with the interviewees. In order to discuss the observed tension with those interviewees who knew the data curation processes most closely, one of my preliminary topics of conversation was the hypothesis of a "trade-off between specificity and generality in data models"; then, I directly discussed events I had observed. In Table 1, I present a census of the data I collected or generated during my fieldwork.

Table 1 - Data generated on site

Empirical effort	
Participant observation	26 weeks full-time office hours
Interviews (avg. duration 60 min.)	30
Other recordings (meetings, conversations)	8
Notes (snapshots, conversations, analytical	665

reflections)	
Meetings (with minutes)	128
E-mail exchanges	1670

Empirical findings

The research site

The business model of *PatientsLikeMe* is centered on commercial research services. These services are fully based on the data that the patient-members routinely collect as part of their self-tracking activities and health community interactions, and revolve around complex work tasks including data aggregation, analysis, and reporting. The clients are organizations from the health care industry, such as pharmaceutical companies or health insurance plans. Through the sale of services *PatientsLikeMe* secures funding for the expensive R&D work that is necessary to develop the system, and for the scientific research that the organization conducts and publishes. A main, overarching concern for the organization is to collect the best possible data, i.e. data that inform, telling us something about a life experience or event that some patient is going through somewhere. Without sufficient amounts of good data to be worked on, the organization could not survive, lacking the raw matter that fuels both services and research efforts.

To the patients, the system represents a possibly easier way to track her health in detail, allowing them to build, over time, a sort of structured journal that stores and summarizes their health life. Most importantly, patients use the network in order to connect with other patients like them. They find support, offer help, find alternative treatment regimes – in the hope for a cure, information about equipment and lifestyle modifications, ask for suggestions or simply communicate their feelings and experience to someone familiar with their experience. This can mean a lot to some patients, such as those who do not feel understood in their life context (e.g. fibromyalgia patients), or those who do not know any experts in their disease, such as the bearers of rare diseases, a relevant portion of the patient population that has perhaps received insufficient attention from medical researchers.[2] To many patients, the site is a place for sharing pain and consolation.

Patients input data on their health status over time, constructing a story of their health life along several dimensions. Through a number of tracking tools, they contribute information regarding the most relevant clinical aspects (e.g. symptoms, treatments, hospitalizations, quality of life) at a time and place of their choice, using the equipment they have and from the context of their daily life. The core dimensions of the patients' health life are captured through the tracking of conditions (and related events e.g. diagnoses, first symptoms), of treatments (and related parameters, e.g. drug dosage and frequency), of symptoms (and related severity), and the eventual relationships between these entities (e.g. a symptom associated with a drug as its side-effect). Other tools capture other health aspects, either generic (e.g. weight) or specific (e.g. lab tests). Without tracking these health dimensions, one could say little about the life experience of the patients.

The system automatically computes scores and charts displaying a longitudinal overview of the medical history of the patients in their individual profiles. Patients can read their profile to try and understand the patterns of their health course. Also, they can browse through a number of report pages that the system automatically creates, on which data from the patient community are globally aggregated in order to provide a snapshot about specific medical entities: there are symptom pages, treatment pages and condition pages, all reporting various descriptive statistics. A symptom report page, for instance, displays statistics of the distribution of severities of the

[2] Estimates suggest that rare diseases affect 300 million people globally. Yet no FDA-approved drugs exist for 95% of rare diseases (RARE, 2014).

symptom,[3] a list of the treatments that patients take for the symptom, and demographics of the patient population currently suffering from the symptom. These pages also host various hyperlinks that link to other patients or medical entities. On the sidebar of a symptom report page, a number of links lead to forum discussions where patients are talking about the symptom, or to the profiles of other patients suffering from the symptom. Page after page, the patients can discover a virtually endless network of relations with other patients and health situations.

Tracking is instrumental to improving patients' socialization opportunities. Scores and charts can be important matters for discussion with other patients. Patients read scores in order to understand their health through an objective, third-person narrative. They tend to welcome with excitement eventual progress in their metrics – hopefully demonstrating actual health progress.[4] Patients are disappointed when they do not see the change they expected, and comment about it with other patients. More importantly, the *PatientsLikeMe* system is more able to connect patients to other patients if they share some piece of data about their own health life – if they track some health aspect. The system is engineered as to compute and display connections and links to other patient profiles, activity or discussions, based on given data points. For instance, the system is able to link patients to the most appropriate forum rooms if they input the condition they suffer from. A host of features – predominantly the dynamically computed links to other patients that are disseminated through the website's many pages and reports – facilitate interaction on the basis of data points that intersect at the convergence of different patient life trajectories. The features through which the *PatientsLikeMe* system draws and structures opportunities, spaces and avenues for social interaction that did not previously exist is a prominent characteristic of this network – one it shares with most prominent social media sites – elsewhere defined as '*computed sociality*'

[3] Symptom severities are captured along a NMMS (none, mild, moderate, severe) scale.
[4] See, Chapter 5 'On tuberculosis and trajectories' in Bowker and Star (1999), for an stimulating discussion on the relationship between health measurements, and biography.

(see Kallinikos and Tempini, forthcoming).

At the other end of the *PatientsLikeMe* system, the research team gathers and analyzes the patient data, to produce scientific evidence of real-world medical phenomena. Exploiting the continuous updatability of Web-based applications, the organization develops, updates, and tweaks the system in order to make it more efficient for the collection of research data.

The problem of patient engagement

The 250,000+ patients in the system[5] come from the most diverse life experiences and contexts. They carry disparate combinations of conditions, symptoms, and other health factors. To cater to all this diversity and to ensure it is adopted, the system needs to be as contextually relevant and flexible as possible. The system's ability to collect data is dependent on its capability to keep the patients engaged in interactive data collection tasks. It needs to motivate patients to come back and continue self-reporting. Engaged patients – regularly visiting the website and participating in its routines – enable longitudinal data collection over time, traditionally a very expensive and valuable research feature. The need to keep patients engaged and inputting data over time characterized much of the effort put into developing the system. It is a big concern, since poorly engaged patients can omit to input very important clinical information.[6] As a researcher at the organization explained,

> '*Right now you* [as a patient] *can load in as many conditions as you want. You might forget to mention the stage-four breast cancer that you survived ten years ago, which clinically is very important, but might not be what you are thinking about right now.*'

[5] As of September 2014.

[6] However, even engaged patients can omit very important information because of self-reporting biases.

Also, the system must be able to allow the reporting of the unexpected, rare medical events that can turn out to be valuable for research purposes – initiating potential discoveries. Rare events can be detected through the engagement of large cohorts of patients and an open data collection process, one that does not constrain data collection to a limited set of possible medical events. An open data collection process, however, needs to be fine-tuned in order to distinguish real evidence from incorrect data. As an executive explained,

> *‘This is a bit of a generalization,* […] *but in the long tail of our data there's probably three things: there's probably patient error, fraud (although I don't think we have a lot of that), and really interesting stuff. And it's hard to figure out which they are […] But there are gems out there…’*

In order to develop data that expresses valuable information – informative data – the system needs to collect as much data as possible. Some meaningful but rare correlations will only emerge out of large numbers. The system needs to be easy to adopt and flexible to suit a patients' context and motivations. However, several factors make such data collection a challenging feat. For starters, it proves to be particularly difficult to have patients input data at the desired intervals – according to a constant time scale – instead of at random times. It also proves to be difficult to have patients complete multiple questionnaires or data collection tasks, which are separate but medically related. Often, patients complete only a partial set of tasks, being interested in tracking only a few of the health dimensions. Partial or temporally distant completion of the data collection task often prevents researchers from reliably relating two data points and conjecturing upon their relationship. Regular and comprehensive data collection would allow attempts to be made to draw a comprehensive picture of patients' health status, but often patient profiles will contain just a few isolated data points. Researchers cannot do much with such patient data. For instance, a reported change in symptom severity would prompt a researcher to control for changes in the treatment regime. In the case of data on the treatment regime being missing, such a hypothesis could not be validated due to a lack of data points. The isolation and consequent lack of context of the data points is one of the most disruptive issues for research conducted through an open and distributed data collection architecture. As one of the managers liked to say, *'No data* [absence of data] *is not "No" data* [data stating 'no'].*'*[7] Still, in order to maximize the data collection chances, the system supports data inputting at any frequency and schedule, as long as a minimum frequency is met.[8]

Increasing information production through local context flexibility

In order to be flexible enough to adapt to patients' life and local context, the system has the built-in capability to customize, to a certain degree, both patient profiles and the underlying data structures representing medical phenomena. At one level, the system is able to personalize profiles, adding custom tracking tools (e.g. lab result tracking tools, condition-specific patient-reported outcome tracking tools), depending on the conditions that the patients report or in response to a request from an individual patient. At another, deeper level, the community of patients shapes the medical representations captured in the data structures. The great majority of the conditions, treatments, and symptoms have been added upon patient request, one at a time. The tracking tools allow patients to log requests for the creation of medical entities or definitions that are not already present in the database. The system has been developed with the aim of recording the patient experience through patients' own definitions, with the conviction that patient experience and language have often been neglected by expert clinical practice. As a *PatientsLikeMe* researcher

[7] This is a form of the popular statement 'the absence of evidence is not evidence of absence' (in this formulation, attributed to the astronomer Sagan; see Wikipedia, 2014).

[8] While the system needs to be flexible, to support different life routines and goals, on particular occasions it constrains access to specific areas of the tracking tools. For example, when a patient does not update her symptom severity scores for more than a predetermined number of days, the system will not allow her to review her symptoms data without first inputting updated symptom severity data. She will also not be able to track a new symptom before providing a new symptom data update. In this way, the system tries to force data inputs when a patient's data inputting falls below a specific threshold, thus obtaining compliance through constraint.

argued when presenting at a major American medical informatics conference, 'the medical profession keeps that [expert] language away from them [the patients]'.

There are reasons for these strategies for the maximization of the system's contextual flexibility. First, such a vast and diverse patient user base implies very different patient experiences in all health dimensions. A major point of differentiation regarding patient experience is conditions. Different conditions mean different patient experience, implications and coping strategies. A flexible architecture shaping the system depending on what information is available about the patient allows the system to respond differently to patients living through very different experiences. For instance, the staff members associate each condition to one of six condition categories.[9] A condition category determines which questionnaire a patient is asked to complete regarding her 'condition history', on a page that attempts to metaphorically take on the function of the clinical interview in traditional patient-clinician encounters. Through this survey, the system asks questions that are appropriate to the nature of the condition. A chronic condition has a very different course and implications from a pregnancy-related condition. Also, depending on the patient's condition, the system selects and associates to her profile specific sets of tracking tools related to the "standard" experience of the disease and its measurement – for instance, patient-reported outcome (PRO) surveys or specific lab result tracking tools.[10]

A second reason for building a flexible system is that patients can have different levels of medical literacy, ranging from doctors to the medically quasi-illiterate. Also very varied is the level of patient understanding of the research scopes for data collection. Despite the organization's

[9] The condition categories, driving different condition history questionnaires, are infections, chronic diseases, pregnancy-related, mental health, events and injuries, and life-changing surgery.

[10] This, however, is possible for only a small number of conditions. Establishing what the standard set of tools should be for a specific condition requires expensive, in-depth research. Therefore, this tends to be accomplished mainly in association with condition-specific, funded research projects.

efforts to make this clear since patients' first landing on the website homepage (a link 'How we make money' explains the business model and mission of the platform) many patients seem to collect data only in fulfillment of a personal journal – with resulting difficult to decipher language. The functional components of the system – electronic forms with concatenations of structured questions, data input interfaces, and data models – are considered instrumental in 'helping to guide the patient to the form that is most likely to be medically accurate', as an informant explained in regard to data collection on drug forms.[11]

Encouraging and guiding patients to complete data collection tasks is a goal that shapes the design of the system. Trying to improve patient engagement often means simplifying things, decreasing the complexity of the technology and, crucially, that of its semantic context. One example of this was the introduction of the 'fuzzy dates' feature, which allows patients to record incomplete dates. The feature was introduced in order to make sure that more patients would input dates in association with medical events. A patient who has lived with a chronic condition for a long time may not remember the exact date of her diagnosis or her first symptoms. Previously, the system required exact dates, constraining patients to fill in all date fields in order to record the data. The organization realized that this design was leading many patients to avoid inputting any dates and thus failing to complete the data entry task. By introducing the possibility of inputting just the year, or just the year and the month, of some events, the system sacrificed data specificity for better patient engagement and more data.

Increasing information production through semantic context

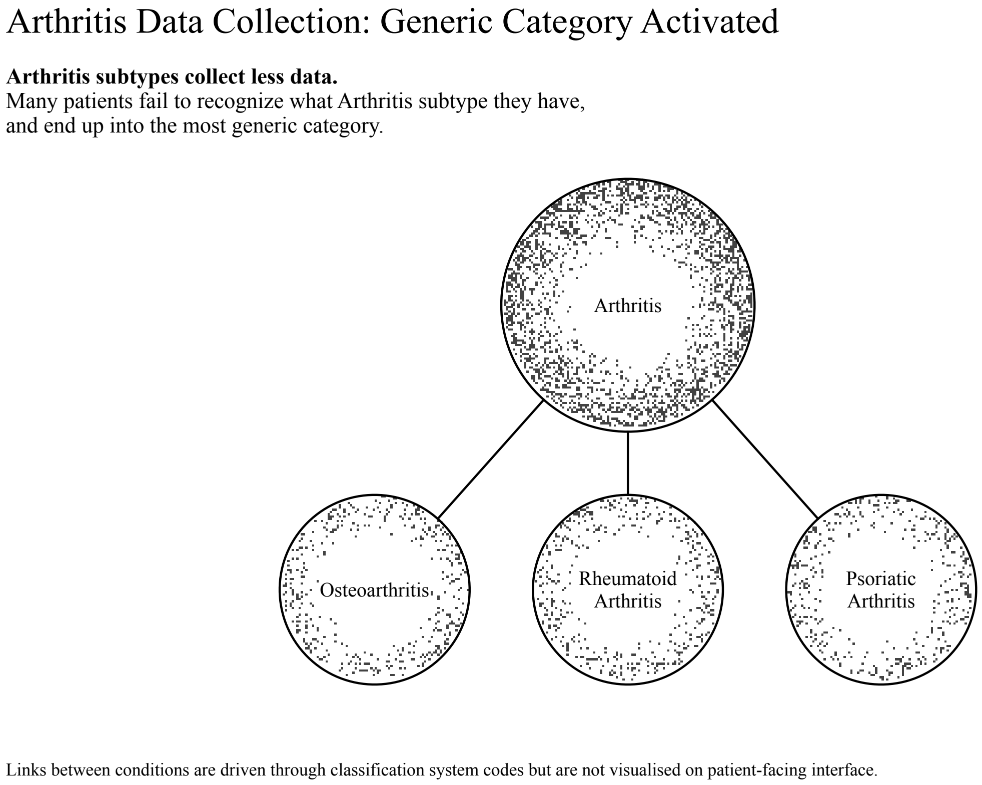
The flexibility to fit local contexts is instrumental for supporting better engagement from patients. Better-engaged patients produce more data. More data increase the informative potential

[11] E.g. free form, pill, vial and etcetera.

of the underlying database. However, the flexibility is sometimes reduced in order to favor other, competing needs of information production. This happens when the priority is to avoid impoverishing the semantic context of the collected data. For instance, the need to differentiate between patients suffering from taxonomically close conditions (subtypes of the same parent condition), but whose lived experiences are actually very different, led the clinical specialists to force patients to select one of the subtypes when as they added a condition to their profile, by disabling the parent condition (disallowing patients from adding the parent condition to their profile). Recall the fictional vignette in the introduction, about arthritis. As a clinical specialist explained,

> 'There are conditions for which there is sort of a colloquial way of talking about it, that doesn't necessarily get at the underlying pathology or the specific kinds of treatments one would need to have in order to develop or understand that condition.'

The generic 'arthritis' was initially a condition that patients could add to their profiles, but it was subsequently deactivated. Many patients were adding 'arthritis' to their profile while actually they suffered from one of its several subtypes. The arthritis subtypes of osteoarthritis, rheumatoid arthritis and psoriatic arthritis, to name a few, involve very different life experiences. After reviewing the data that they had collected over time, and finding that too often patients were adding the generic 'arthritis', the staff decided to require patients to choose the subtype of their condition. Once the generic 'arthritis' had been deactivated, patients could no more add the parent condition to their profile. Patients were constrained to either find the name of their condition in a better-specified form (a subtype definition), or else not add the condition to their profile. The newer data structure, making a distinction between subtypes of arthritis, required from patients data reporting at a higher level of specificity, and better differentiated between patients and their respective experiences. In this case, semantic context was increased at the expense of patient engagement (and in turn data scale).[12] Figures 1 and 2 descriptively represent this trade-off in a simplified fashion, by showing two alternative set-ups of condition categories and the consequent effects of the scale of data collection.



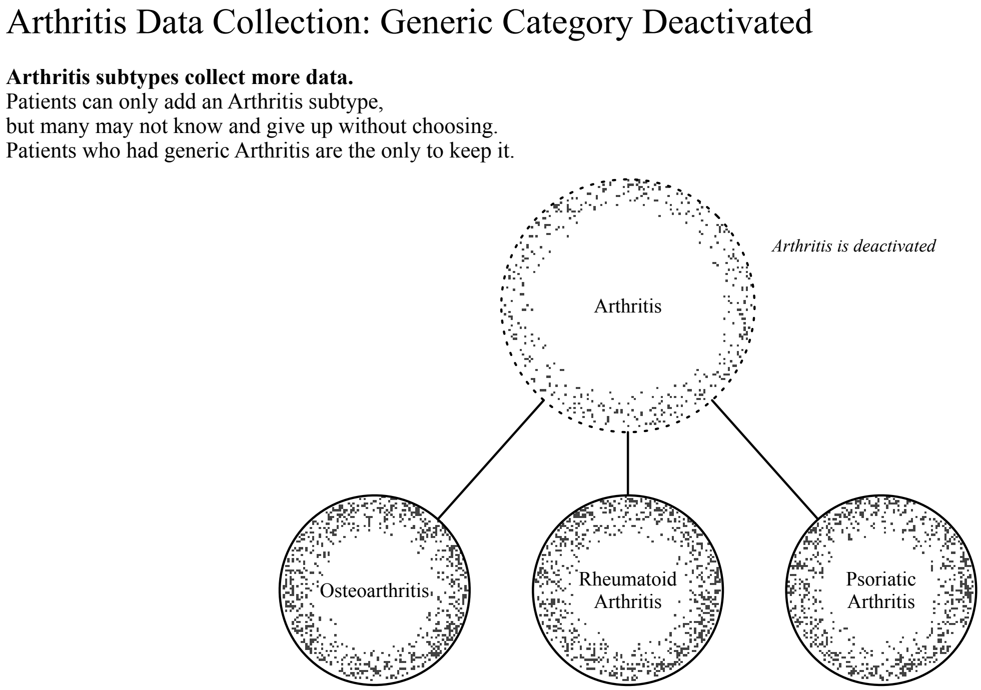
Arthritis Data Collection: Generic Category Activated

Arthritis subtypes collect less data.
Many patients fail to recognize what Arthritis subtype they have, and end up into the most generic category.

Links between conditions are driven through classification system codes but are not visualised on patient-facing interface.

Figure 1: Data collection including generic Arthritis

[12] Obviously, there are simpler conditions where it would not make sense to split the world in two. For example, it would be detrimental to divide patients into those with a ‘broken right leg’ and those with a ‘broken left leg’; aggregated data provides sufficient power in this case. The same is true, but for different reasons, with generic conditions of which patients rarely get to know the type (think ‘flu’).



Arthritis Data Collection: Generic Category Deactivated

Arthritis subtypes collect more data.
Patients can only add an Arthritis subtype,
but many may not know and give up without choosing.
Patients who had generic Arthritis are the only to keep it.

Figure 2: Data collection excluding Arthritis

As the organization tailors the system in order to produce more information, both data scale and semantic context shape system development efforts. Obviously, the organization makes use of various metrics and analytics to support meetings and decision making. During the observation period, the staff often discussed how to gauge the information potential – the potential to produce information – lying in the database. The best metric available, though the staff did not find it entirely satisfactory, was called the 'patient-outcome years'. An executive described this as 'the ultimate, the toughest measure of the value of the fundamental underlying database.' Without going into the complexity of its calculation, the metric aimed to estimate the information potential captured in the database as a product of the volume and density of rich data (specifically, patient-reported outcome data). For the purpose of this paper, it should suffice to say that the information potential of the data was perceived to depend on both the specificity and the scale of the data. An executive explained,

> 'The [patient] outcome years is sort of the last measure; it's this, sort of, if you have active users who are engaged, and are giving data over time, [it] measures how well they're contributing to the fundamental database.'

Local versus semantic context in user-generated data collection

The struggle between the conflicting demands for local context flexibility and data specificity richness played out in a more complex way in another feature of the system. As I have explained early on, the system is designed to allow patients to track a number of medical entities, including treatments and symptoms. Here, I analyze the example of symptom tracking. The symptom-tracking tool is a standard tool that all patient profiles have. Patients track a list of symptoms, recording for each of them severity scores and two types of associations with treatments – a treatment can cause a side-effect symptom, or a symptom can be the reason for taking a treatment.

The system automatically adds symptoms to the tracked symptom list on the patients' profile in two ways. First, upon account creation, the patients profile is attributed five generic symptoms deemed applicable to any patient experience.[13] Second, the system automatically adds a number of condition-specific symptoms to the patients' profile when the patients add a condition to their profile.[14] Through the attribution of specific symptoms to profile of patients suffering from a determinate condition, the system is able to demarcate a minimum common denominator of the patient experience. All patient profiles can then be juxtaposed and compared based on this set of

[13] The five generic symptoms are anxious mood, depressed mood, fatigue, insomnia, and pain.

[14] This feature, however, is limited to the minority of conditions about which the staff has had the opportunity – usually in the context of funded commercial research projects – to carry out the research required to infer the symptoms most characteristic of a patient's experience of the condition.

shared symptoms. Patients have been found to track condition-specific symptoms quite variably, however, probably because it is burdensome to repeatedly track several symptoms some of which one might even not experience. Patients can also edit the tracked symptom list on their profile, adding symptoms as they wish, by clicking on links on the symptom report pages or through the search feature. In this way, patients can customize their profile and tailor the symptom list to their own patient experience. If they are unable to find a matching symptom through navigation or the search feature, they can issue a request for the creation of a new symptom, providing a patient-generated definition of it. Patients had added, by request and one instance at a time, nearly all of the roughly 7,000 symptom categories that were being tracked by the website at the time of my fieldwork.

Often, the symptom that patients are experiencing and want to add to their profile is already represented in the database. There are a number of reasons why patients might be unable to recognize their experience in an existing record. Impatience in reviewing search results, or misspellings that the spell-corrector fails to pick up, are just two of the potential reasons for a redundant symptom creation request. Most importantly, unconventional, folk, and patient-generated definitions might not match easily with the existing record. For these or other reasons, if the matching is not successful the patients can submit a request for the creation of a new symptom record.[15] The staff reviews new symptom requests. A team of clinical informatics specialists manages the incoming new symptoms from a dashboard in a restricted-access area of the website. The staff members perform a number of tasks as part of the request-review routine. First, they research the database to verify that the symptom is not already present in the database. They also search medical resources (UMLS, PubMed, E-Medicine portals, Wikipedia, Google) to

[15] This is also possible for other medical entities such as conditions and treatments.

investigate whether the definition provided by the patients does in fact describe a symptom.[16] They keep in communication with the patients, explaining the status of the review and often asking for clarification or further information. In a short series of written exchanges, the patients can explain their experience further to the staff, participating in the investigation to understand and define the clinical situation at hand. Sometimes the patients might be describing a symptom that is already represented in the system, only in a different language. Often, the patient definitions are more specific under some aspect (e.g. laterality, or emotional nuance) than the description given by the expert terminology.

Storing more specific symptom definitions in patient language generates more information – increasing the power to differentiate between two different patient experiences – while increasing the system's flexibility to deal with local contexts, as long as different patient-generated definitions can be related to each other or to a common root phenomenon. An unrestrained capability to create symptoms is not, hence, intrinsically desirable for research. Pursuing differentiation through such an open, participatory architecture exacerbates a particular challenge. Storing two very similar patient symptom definitions that differ only minimally favors database fragmentation, potentially impeding the aggregation of similar cases at the level of granularity that is relevant for research purposes. The inability to equate and aggregate data related to similar symptoms can hamper the validation of a research hypothesis.

Once the staff members believe to identify the clinical situation described by the patients, they can take a number of actions on the symptom request. On the one hand, they can refuse to create a new symptom record and merge the patients' symptom definition into an already existing

[16] The ontological status of certain medical entities is often disputed, e.g. in the case of syndromes. Sometimes the boundary between symptom and condition is blurred and shifting. Simpler cases can be dealt with more straightforwardly, for instance when the patient has entered an entity that is clearly not a symptom, e.g. a drug.

symptom record. The patients' symptom data is thus aggregated with other patient data linked to this symptom. Such decisions are not always welcome by the patients and may strain their engagement with the platform, leading them to stop actively collaborating, to become inactive or to ask for the deletion of their data. For this reason the staff members try to explain and include the patients in the symptom review process. On the other hand, if the review is concluded positively, the staff members approve the new symptom and fill a symptom configuration form in the restricted area of the website. The configuration form stores the essential information about the symptom, including a textual description and codes to link the new symptom category to expert terminologies such as SNOMED, ICD10, ICF, and MedDRA LLT. Other actions that staff members can take on a symptom request include archiving it, when a sound decision cannot be reached, or splitting it in more symptoms, when the patients have erroneously inputted two or more symptoms in the same string.[17]

Through this open, participatory data collection process that recognizes the patient a role of observer and operator (see also Kallinikos and Tempini, forthcoming), the system is able to detect and capture new entities into symptom categories. Under the category of symptoms, the system hosts two categories of medical entities, symptoms and signs.[18] Symptoms data collection requires flexibility towards patient observations, since symptoms are inseparable from subjective experience. Patients can be very meticulous in differentiating between experiences and sensations, and different levels of literacy and of commitment to the research aspect of self-tracking also affect

[17] For instance, 'toothache cognitive impairment' is a string that can be split into two symptoms 'toothache' and 'cognitive impairment', which can then be added to the database.

[18] Briefly, the difference between signs and symptoms lies mainly in who is able to observe the phenomenon in question. Scheuermann and colleagues define a sign as a 'bodily feature of a patient that is observed in a physical examination and is deemed by the clinician to be of clinical significance' (Scheuermann *et al.*, 2009:119). For instance, a lump can be a sign: both the clinician and the patient can easily observe it. A symptom is instead defined as 'a bodily feature of a patient that is observed by the patient and is hypothesized by the patient to be a realization of a disease' (Scheuermann *et al.*, 2009:119). For instance, the clinician does not directly observe a symptom such as a headache. Only the patient has access to the phenomenon.

the way symptoms are categorized. In its early days, the platform hosted a community for only one condition, Amyotrophic Lateral Sclerosis (ALS), and allowed the tracking of a widely used, fixed list of 40 symptoms developed by clinical experts in the disease. The list captured the most common symptoms in the ALS patient experience as understood by the scientific community. However, managing a social media platform connecting thousands of patients across the globe, it quickly became clear to the *PatientsLikeMe* developers that many more symptoms, experiences, and circumstances characterize an individual ALS patient experience. Importantly, many patients develop co-morbidities, and a platform designed for scientific discovery should be able to capture all relevant patterns.

The patient experience had to be captured more holistically. Open and participatory symptom data collection features such as those I have described were added to the system then. In a following study, Arnott-Smith and Wicks (2008) analyzed the 376 symptom terms that had been created by patients until then and found that 43% of the symptoms could be matched to terms in the UMLS (Unified Medical Language System) meta-thesaurus. However, only 38% of the patient-submitted symptom categories corresponded to symptoms or signs in the UMLS, with other semantic types represented in the symptom data being disease or syndrome; finding; pathologic function; mental/behavioral dysfunction; and body part, organ, or organ component (Arnott-Smith and Wicks, 2008).[19] Other kinds of anomalies, however, are less straightforward to address. These occur when patients input, as symptom entries, complex constructs such as fragments or phrases, multiple clinical concepts, temporal associations, and slang (Arnott-Smith and Wicks, 2008). Also,

[19] Importantly, patients were actually recording co-morbid conditions in 25% of these cases. A cause of this was that the system could associate only one condition with each patient profile. As many chronic patients live with co-morbidities, they were working around this system limitation by storing co-morbidities as symptoms. When, in 2011, the system was developed to allow patients to add multiple conditions to their profile, it became better able to correctly guide this kind of data inflow. The development of a considerably more complex system, in which a patient could associate to her profile any possible combination of conditions, successfully controlled this instance of data collection creep.

and importantly, the researchers found that many symptom terms express 'either a problem or a body part in more granular terms than the UMLS "knows"' (Arnott-Smith and Wicks, 2008: p. 685). Over time, the open and participatory process of differentiation between lived experience and recorded symptom definitions can produce redundancy and hamper the aggregation of data. If patients distinguish between two different types of pain that do not, however, make a difference to medical research requirements, the platform loses informative potential unless it is able to aggregate the data and compute them as instances of the same phenomenon. The flexibility the system needs to adapt with diverse local contexts ends up undermining the systematic and largely automated collection of informative data. A flat, endlessly fragmented data structure, unable to draw existing similarities between symptoms, is collecting data with poor semantic context.

To obviate to the developing situation the *PatientsLikeMe* developers rolled out software features that allowed the staff, in the restricted-access area of the website, to map the patient-generated symptom categories to expert classifications in hierarchically structured terminologies (i.e. SNOMED, ICD10, ICF and MedDRA LLT). Mapping symptom categories to hierarchical terminologies enabled the organization to translate and aggregate related yet different patient symptom definitions when it became necessary for research purposes. This labor-intensive mapping operation – requiring research into the nature of many symptom phenomena – reconstructs the semantic context lost by allowing open, participatory differentiation of patient experience. As a member of staff explained,

> 'There's probably about twenty different ways that people can express pain: nerve pain, bone pain, all sorts of different types of pain. Now, in my back-end view I can see all those ways. […] If someone puts in "red prickly rash on my leg", if there's a specific symptom [that matches this], I can see how that relates to every other person who has had a symptom that hit on the same MedDRA

> constellation [coded against the same MedDRA code]. So, maybe the overarching one is "rash", but you get down to the one [symptom definition] that the patient actually told us about in their own words… it's still gonna bubble up [the patient definition is still going to be represented]'.

For example, symptoms of anxiety are distributed across a large number of different patient definitions. Mapped to the same ICD10 and ICF codes as 'anxiety with telephone' – respectively, F40.2 'Specific (isolated) phobias' and b1522 'Range of emotion' – are symptoms such as 'needle anxiety', 'fear of confined spaces', 'fear of cold (cheimatophobia)', 'fear of heights (acrophobia)', 'paruresis', 'fear of large oversized objects (megalophobia)' and 'fear of work (ergophobia)'. An admin user can easily navigate this constellation of symptoms, grouping them by the same classification code. Constructing a symptom database that can be nested within an existent, expert hierarchy allows *PatientsLikeMe* researchers to aggregate patient data in bigger data pools. At the same time, and on a systematic basis, it still allows the researchers to divide between experiences and the patients that lived through them at a further level of granularity than the existing terminologies allow.

Discussion

In the introduction I posited that in order to understand how organizations developing social media networks exploit, open, distributed, and data-based networking arrangements with the aim of producing information and knowledge, we need to study the processes of data making, and data sense making within the organization. The premise was that social media are systems embedding complex data structures that shape data sense making and information production and hence, in turn, the way the social media infrastructure is governed. In this respect the empirical evidence compellingly shows us that something specific is at play when an organization tries to engage the general public in information production. In the first instance, we observe that organizational efforts to cultivate the information potential of the data are often torn between conflicting demands. These are the demands for local context flexibility and semantic context. A highly engaged patient user-base generates more data, increasing the information potential of the data by increasing its scale in terms of both sample size and longitude. To achieve higher levels of engagement, the system needs to be able to adapt to many specific local contexts and patient experiences, in all their extreme diversity. It needs to be easy to use and customizable. However, we observed that developing the system for higher engagement often reduces the semantic context of the data. The data contain less information, and are less able to show differences and relatedness between phenomena. The system collects more data but these data are, taken individually, less meaningful. Conversely, higher semantic context increases the information potential of the data through the power to differentiate and associate phenomena more finely. To increase the semantic context of the collected data, both the amount of structure and the specificity of the data models need to be increased. However, we observe that more specific or structured data often implies a more constrained and restrictive user experience, with consequently lower levels of patient engagement. The system collects more meaningful data but these data are, in total, fewer.

The complexity of the tasks involved in governing the *PatientsLikeMe* data collection architecture led the organization to take a contingency-based, iterative approach, taking development decisions based on continuous review of the status of the collected 'data pool' (Aaltonen and Tempini, 2014). At times (e.g. fuzzy dates), collecting sufficient relatively vague data was prioritized over collecting precise data in small quantities. Requiring patients to input the exact dates of events long past seemed to prevent some patients from recording data at all. Conversely, in other situations (e.g. arthritis subtypes), collecting more specific data of a certain kind was prioritized over input volume. Forcing patients to choose between arthritis subtypes, at the cost of turning some away, was felt to be the better choice. It is important to note that the value of the collected data was reviewed by considering the informative potential of the whole data pool (Aaltonen and Tempini, 2014). A different informative capacity of the data emerges when the data are treated as a whole rather than individually.

Mechanisms of information cultivation

Information cultivation is the concept that I introduce in this paper with the aim of capturing the strategic, operative horizon in which the daily activities of social media systems development take shape – including gauging the informative potential of the collected data. In order to further explain the evolutions of the *PatientsLikeMe* data collection system that we have observed, I theorize about two mechanisms of information cultivation. First, in the development efforts intended to cultivate information through better patient engagement, we observe a mechanism of data pool extension. Some changes in the system afforded an increased flexibility to adapt to local contexts, which was associated with higher engagement levels. The system could then gather more data from otherwise passive patients (an increase in active population), but also more data from already active patients (and increase in data points density). The data pool could be shaped along two dimensions, hence the choice of the surface metaphor 'extension'. Second, in the efforts to cultivate information through higher specificity and more structure in the data, the active mechanism is one of data pool enrichment. Some changes made data models more precise in differentiating between (and consequently associating) phenomena. Similar phenomena, that otherwise would have been represented as the same phenomenon, were now recorded as different. The movement is one whereby more phenomena diverge, centering upon different data representations. The segmentations and splits that data structures effect on the world are more granular, have a higher resolution. The network of their relationships is more complex and closely interwoven, it is of a richer thread, hence 'enrichment'.

It is important to observe that the mechanisms of information cultivation – data pool extension and data pool enrichment – often have a paradoxical relationship. As shown through the empirical evidence (e.g. fuzzy dates and generic arthritis), both mechanisms increased the information potential of the data by strengthening one of the two factors of information production – scale and specificity – while at the same time constraining the other factor and thereby introducing a countervailing effect.

Over time, the social media infrastructure was developed in a stepwise fashion, with both mechanisms activating at different phases. In the example of symptom data collection, *PatientsLikeMe* developed the feature of allowing patients to enter new patient-generated symptoms, a development from the initial stage of a fixed list of symptoms for a limited number of conditions. Patients could then store more information about more phenomena, capturing new aspects of their lives and experiences. However, as we have seen, the semantic context of the data was unsatisfactory because of the flat structure of the symptom categories. Redundancies and errors among the symptom categories abounded. A second evolution, building on top of the previous, was the introduction of background coding, afforded by new and more powerful database editing tools for clinical specialists. Background coding is a labor-intensive task, often requiring iterative communication between the staff members and the patients. Coding patient symptom definitions to link them to expert terminologies provided the system with the capability to group and aggregate symptoms as needed for research. This feature required more active management of the patient-generated categories by hand of the staff members – as we have seen, sometimes at the expense of the relationship with patients due to disagreement over staff decisions over symptom requests. To summarize the argument, I depict these three empirical episodes – fuzzy dates, generic arthritis, and patient symptom definitions – in the simplifying charts of Figure 3. In the diagram in

Figure 4, I summarize the relationships between the theorized mechanisms of information cultivation, data pool enrichment and data pool extension, and the concepts, on which the theory is built, of semantic context and engagement level.

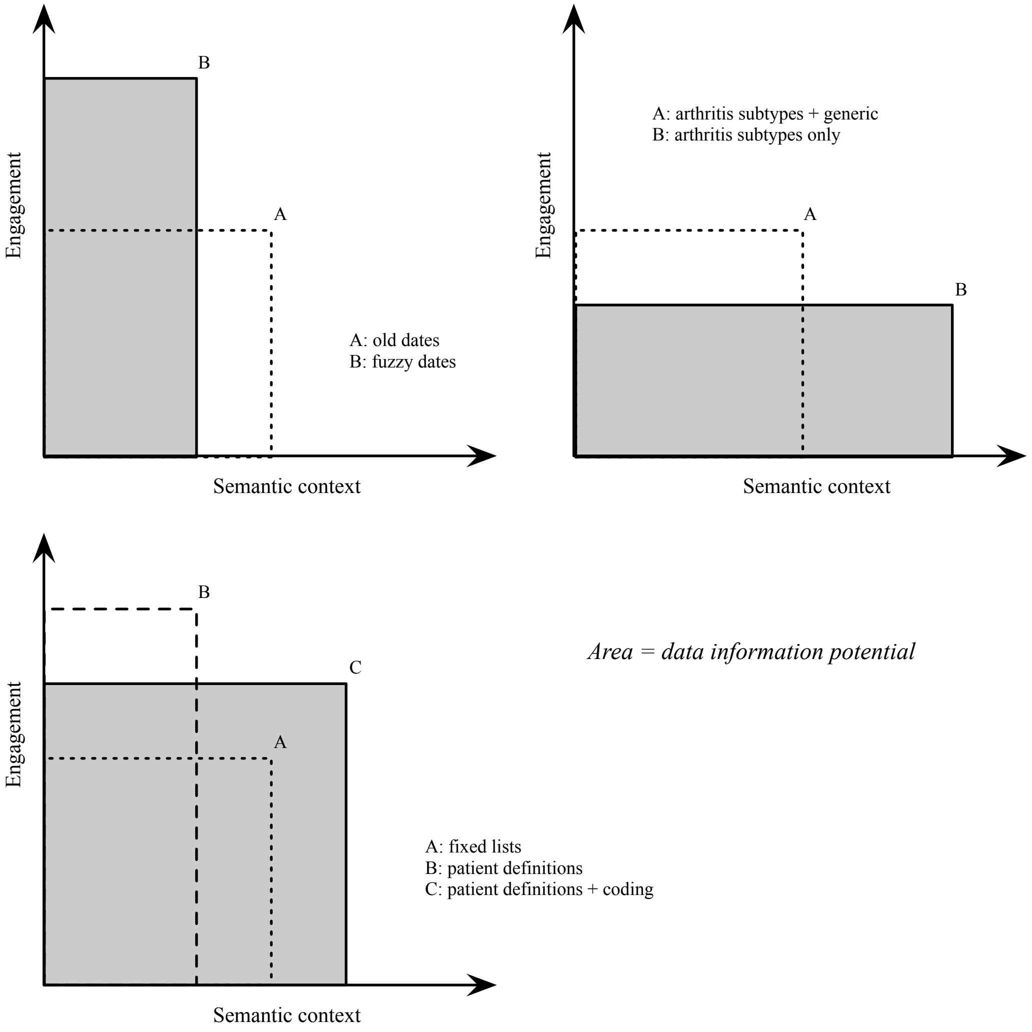


Figure 3: Shifts in information potential, in the examples of fuzzy dates, generic arthritis, and patient-generated symptom definitions

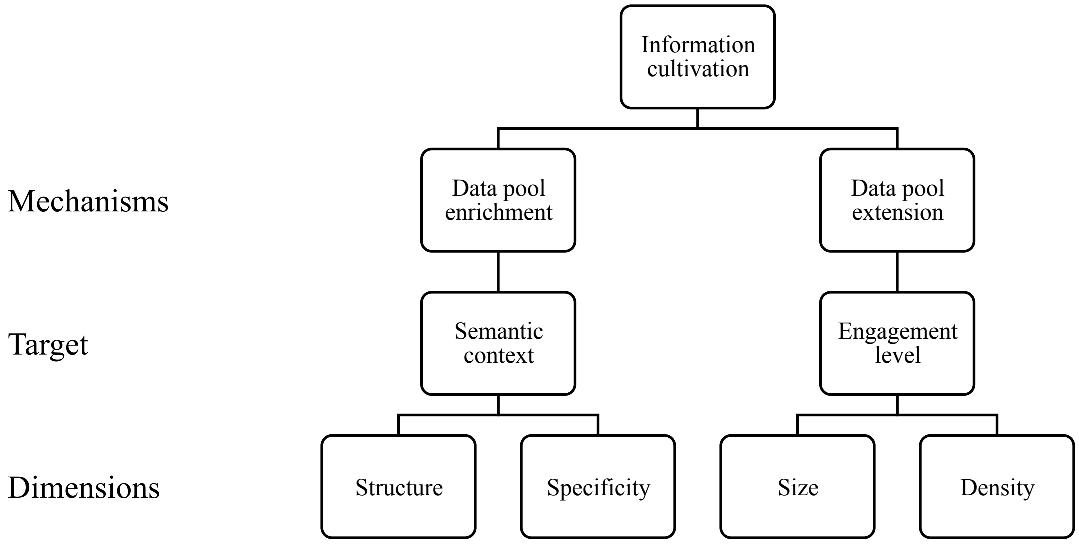


Figure 4: Information cultivation and its mechanisms

PatientsLikeMe and knowledge making in the age of social data

In order to see the relevance of the *PatientsLikeMe* case and the explanatory power of the analytical devices I theorized – the overarching strategy of information cultivation and its two mechanisms, data pool extension and data pool enrichment – we need to situate the organization and the kind of scientific enterprise it encapsulates against a broader background than the crucial but relatively specific context of the use of social media in medical research. As noted earlier, *PatientsLikeMe* should be contrasted to other social media- and research-based organizations on the grounds that its innovative approach to research data collection and clinical discovery is centered on an open, purely distributed and data-based information production infrastructure.

The network is *open* because, through a specific information production architecture, the system allows unknown events and forms of human experience to be captured in a database. First, the immediate availability of the system to anyone that has access to now basic computing and networking facilities allows unknown individuals to make themselves known and report medical data from their own local context (see also Prainsack, 2014). Second, the relatively simple software interface and embedded patterns for data self-reporting allow instances of particular medical phenomena to be reported and made known to the system by such individuals. Third, the flexible architecture for the management of medical knowledge representations allows the recording of unexpected phenomena, whereby instances of unknown identity (i.e. new patient-generated symptoms) are made known to the system and recorded. The system does not impose a strict cognitive grid of phenomenic possibilities. It captures events comprehensively and deeply – as its discovery potential depends on detecting the "long tail" of phenomena that might produce medical breakthroughs.

Second, this information production arrangement is also *purely distributed* because data are contributed by an undefined multitude of patients, from any kind of life context affording basic connectivity and none of which is at any time physically accessible to the researchers in the organization. The only source that the organization has to find out about the patients – here collaborators upon which the organizing depends (see also Kallinikos and Tempini, forthcoming) – and their health lives is the web-based, distributed platform. This aspect perhaps more than others sets the case apart from previous studies of development of data structures in the context of distributed science, where projects seem to involve multiple but knowable and finite contexts and operators (e.g. Millerand and Bowker, 2009; Ribes and Bowker, 2009; Ribes and Jackson, 2013). Finally, the information production arrangement in *PatientsLikeMe* is also essentially *data-based*, because the inaccessibility of the patients and their life contexts makes the descriptions, labels, categories, scores, aggregates, and counts that the system stores and computes the only material at the center of the research work.

One broader domain to which the *information cultivation* challenges identified in this case should be associated is that of those organizations that critically depend on their ability to leverage social media technologies for the production of information through undefined, ephemeral, and distributed relationships with the members of the massive publics they serve (Mathiassen and Sorensen, 2008; van Dijck, 2013). This broader domain includes social media organizations but also overlaps with the ostensible development of "Big Data". A distinctive feature of these innovative data-based, or data-intensive, organizational forms stands in the nature of the relationship with their technological underpinnings – which are not only tools of transformation of work into information processing and 'reading' (Kallinikos, 1999; Zuboff, 1988) but also the raw matter that is needed for the construction of new products and objects derived from digital data. One common denominator across the colorful range of entrepreneurial efforts of these initiatives seems to be the assumption that data can always be variably and indefinitely repurposed – the meanings of data being largely independent from the purposes for which they are generated. The data social media users generate while going about their everyday lives are looked at almost as an open journal displaying their needs, thoughts, concerns, and tastes (Gerlitz and Helmond, 2013; Kallinikos and Tempini, 2011). In the age of Big Data, some argue that virtually any kind of digital trace, if provided in enough quantity, has the potential to unearth surprising discoveries (boyd and Crawford, 2012; Mayer-Schönberger and Cukier, 2013). No doubt these socio-technical developments will generate great value, and unforeseen social or personal gains in many domains. However, what the evidence from the *PatientsLikeMe* case seems to suggest is that the production of (scientific) information from social data collected through social media is characterized by specific information infrastructure development challenges that shape and are shaped by the specific and to some degree contingent socio-technical configuration of people and systems that such initiatives bring about.

Governing through social denomination

In a social media network such as *PatientsLikeMe*, data structures are developed to adapt to the contingencies of data collection in an open and distributed setting. The staff develops the system and its embedded medical knowledge representations in reaction to the evolving outcomes of the data collection arrangement, which keeps the patients and data structures woven together, inseparable in the data thus produced. The very configuration of this scientific arrangement shapes, in the specific ways I have defined, the kind of medical evidence, and in turn knowledge, that is produced. Social media technology and data structures are not neutral research partners, in terms of how much they allow to do or to know about patients and their life contexts. In a social media network, it is crucial to elicit desired levels of data-generating user engagement. Developers need to enable the patients to tailor the systems to their experiential context. The data collection must remain sensitive to the diversity of medical phenomena, and the patient language in which they might be reported.

Blindly imposing constrictive data collection frameworks might be lethal for the scientific enterprise. As Bateson explained, conclusive, pre-emptive framing of phenomena destroys the possibility of learning (Bateson, 1972; Kallinikos, 1993). The system needs to be able to adapt, as it is upon its capacity of supporting the patients' statements of a difference in experience that depends its own adoption in the patients' own sense-making of their health situation. But, as we have seen in the example of the symptom data reporting, the data pool fragmentation that uncontrolled proliferation of patient-generated data categories could give rise to would not make the information production enterprise viable. The organization needed to develop reporting architectures that allow similarities between phenomena to be recorded, and data on similar phenomena to be aggregated, for successful scientific research to reliably take place.

The mapping of patient-generated symptom definitions through expert classification codes allows the system to traverse the patient language and aggregate symptoms that medical researchers might not need to separate for their own research purposes. The operation aims at reconstructing the meaning to the symptom definitions, that would otherwise get lost, which arises by putting a definition in relation to other symptom definitions. In a double-sided movement, the meaning of each symptom definition is strengthened by the opposition to the other definitions, which are not same (Bateson, 1972; Jacob, 2004; Kallinikos, 1993), but also by the recovery of the eventual overlaps of a category's semantic field to others, which allows to draw, by gradients of difference, the network of relations of a symptom definition with all the others.

The paradoxical tensions of information cultivation, where an organization needs to govern the user base of its social media network at one time to enable and constrain, guide and follow, differentiate and overlap, are of paramount importance for understanding social media. Through the fine-tuning of data structures, a social media organization tinkers with the denominators of social events and phenomena (Bowker and Star, 1999), according to its information production imperatives. In the context of an open, distributed data collection network, what I define as 'social denomination' makes possible not only to pinpoint and compare but also to access, survey and, most importantly, aggregation and computation of otherwise inaccessible contexts. Social denomination defines the situation, in the management of a social media network, where parties are involved in the definition of minimum common denominators that make social (medical) objects manipulable, countable and represented. Boundaries between medical entities such as conditions and symptoms, or coordinates of events such as diagnosis dates, are continuously shifted according to information production goals. By loosening the requirements for a reported diagnosis date, by requiring all arthritis patients to specify the subtype of condition from which they suffer, and by reviewing patient symptom definitions, the organization behind *PatientsLikeMe* is involved in denominating social objects, configuring the lines of convergence along which patient experiences are made to become same (Bowker and Star, 1999). Far from being an original development and tracking back to the origins of taxonomy and statistics (Rose 1999), social denomination operations acquire however a particular importance in social media because they are conducted frequently, often repeatedly, and on a continuous basis, drawing and re-drawing the boundaries of objects or subjects at each take (Abbott, 1988). The sensitivity of these operations in realms such as medicine is obviously paramount as shifting boundaries defining phenomena can make the difference between normal and pathological, and the practical consequences that might follow in terms of personal health management and health care (Lowy, 2011).

The importance of this development is not negligible. It not only shapes at a fast rate the scientific evidence that is produced, and the boundary and identity of social objects and subjects but also reconfigures the multiple data associations that allow constructing webs of links to connect patients to each other. A symptom report page, for instance, dynamically displays a host of links to relevant treatments or affected patients, drawing socialization trajectories and connecting a patient to other virtual spaces (e.g. forum rooms) or patient profiles (for a more in-depth discussion, see Kallinikos and Tempini, forthcoming). Social denomination is foundational for the form of '*computed sociality*' (Kallinikos and Tempini, forthcoming) that the social media infrastructure constructs, and the overarching technique through which a virtual community – such as one gathered and shaped through the *PatientsLikeMe* platform – is governed.

Conclusion

In *PatientsLikeMe* medical research involves delving and sifting through great amounts of data. Researchers browse through the vast database, their research context being labels and numbers of events, patients, conditions, drugs, and symptoms. Within this cognitive environment, scientists inspect and traverse the database in multiple ways, selecting and extracting meaningful patterns out of a mass of decontextualized data (Aaltonen and Tempini, 2014; Kallinikos, 1993). In digital data, patient life trajectories (Bowker and Star, 1999) can be deduced, juxtaposed, and represented in data constellations (around specific medical entities, or data points; displayed in report pages, profiles, or search results), abstracted from the space and time in which those trajectories unfolded. The data pool is a relatively smooth and homogeneous cognitive environment, far removed from the complex real world to which it refers (Borgmann, 1999, 2010).

However, behind the malleable data structures and data pools there is a world in constant movement, which, as we have seen, is able to strike back against pre-emptive attempts (Latour, 2000). The development of a social media infrastructure aims to address real-world conditions affecting data collection (here patient concerns, engagement, motivations, literacy, health status, life context) that, however, remain for the most part unexpressed in the data (Bowker, 2013). This is only in part an epistemological issue (Heidegger, 1962; Wittgenstein, 1953). There is more to this phenomenon than the inevitable limitations of the distributed application of standard analytical reductions. Patients perform data collection for purposes and with hopes that remain unspoken and are different from the purposes of the researchers cultivating the database. They participate in the network not only to participate in research, but also to find a cure and, mostly, to socialize with other patients; they are looking for empathy, solidarity, a potential cure, or simply coping strategies. Multiple and unexpressed perspectives are finding confluence in social media, shaping the collected data.

In this light, I would like to recall the tweaking of the arthritis condition categories episode,[20] which shows how organization and patients had different ideas of what is a meaningful distinction between two arthritis patients. For arthritic patients, coping strategies might be the main concern. To alleviate painful everyday experiences would mean success. From their

[20] Whereby the generic 'arthritis' form was disabled, requiring patients to choose a subtype. This episode saw the organization moving the boundaries defining arthritis conditions, and consequently reshaping the patient groups and sociality created through aggregation.

perspective, there might be not much difference between themselves and patients of another arthritis subtype. However, the patients shape also the space in which the research efforts unfold, when they input data in ways that make sense for themselves or for their fellow patients. They are a gateway to an experiential context that the researchers cannot reach in any other way. In the arthritis case, it became necessary to improve the informative potential of the database by dividing arthritis patients into smaller, more granular groups – the perspective of the researcher being that the biological mechanisms underlying the experiences of different arthritis subtypes might well be different.

A birds' eye view of what we observe throughout this case is that, as social media networks come to embrace society with unprecedented breadth, the social and information are increasingly founded upon each other. Social interactions are intermediated by more and more complex data structures so that they systematically produce more information. At the same time, data structures and information are increasingly shaped by broader and broader social contexts (e.g. patient symptom definitions) – bringing into focus social denomination and its struggles. The paper concludes before opening a topic that clearly was beyond scope of the current research goal. Understanding these consequences of social media technologies for practices and politics of research and health management is something that has remained at the edge of this paper and which I have only sketched, concerned as I was in establishing the detailed empirics, and associated theoretical tools, that could inform and shape more research to come.

In this article, I have presented a study of a social media network through a particular research perspective, documenting the efforts of the owner organization as it has tried to improve its capability to produce information from the data users generate. I have theorized the concept of *information cultivation,* the *data pool extension* and *data pool enrichment* mechanisms and the technique of *social denomination* with the hope that they can help us to understand the specific challenges characterizing such an enterprise. This article has hopefully raised many more questions than it helps to answer. Many other questions could and perhaps should have been asked, however, my assumption throughout has been that social science needs to lay detailed empirical foundations before embarking on discussions of a more critical, ethical, or normative character.

Acknowledgements

I am very thankful to the editor, the reviewers and the special issue guest editors Hamid Ekbia, Jannis Kallinikos and Bonnie Nardi, who patiently encouraged me to take the argument further.

This research was developed as part of my PhD project and it would have not been published without the support of numerous friends and colleagues at LSE, KCL and the wider academic community. I am especially indebted to Jannis Kallinikos, Barbara Prainsack, Michael Barzelay, Nikolas Rose, Carsten Sorensen, Ola Henfridsson and Sue Newell for their support and feedback. Early or later drafts have also enjoyed comments from Jochen Runde, Aleksi Aaltonen, Des Fitzgerald, Ilana Löwy, Cristina Alaimo, Paul Leonardi, Attila Marton, Silvia Camporesi, Tara Mahfoud and the ICIS 2013 DC friends Mohammad Moeini, Koteswara Ivaturi, Brian Dunn, Ina Sebastian, Inchan Kim, John Qi Dong, and Liliana Lopez. Thanks to all of you.

References

Aaltonen, A. and N. Tempini 2014. Everything counts in large amounts: a critical realist case study on data-based production. *Journal of Information Technology* 29(4): 97-110.

Abbott, A. 1988. Things of Boundaries. *Social Research* 62 (4): 857–82.

Agre, P. E. 1992. Formalization as a Social Project. *Quarterly Newsletter of the Laboratory of Comparative Human Cognition* 14 (1): 25–27.

Arnott-Smith, C. and P. Wicks. 2008. PatientsLikeMe: Consumer Health Vocabulary as a Folksonomy. In *AMIA Annual Symposium Proceedings* 2008, pp. 682–686. Available at: http://www.ncbi.nlm.nih.gov.proxyiub.uits.iu.edu/pmc/articles/PMC2656083/pdf/amia-0682-s2008.pdf (accessed November 23, 2014).

Bateson, G. 1972. *Steps to an Ecology of Mind.* London: University of Chicago Press.

Berg, M. and S. Timmermans. 2000. Orders and Their Others: On the Constitution of Universalities in Medical Work. *Configurations* 8(1): 31–61.

Borgmann, A. 1999. *Holding On to Reality: The Nature of Information at the Turn of the Millennium*. Chicago: The University of Chicago Press.

Borgmann, A. 2010. Orientation in Technological Space. *First Monday* 15(6): online. Available at: http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/3037/2568 (accessed November 23, 2014)

Bowker, G. C. 2013. Data Flakes: An Afterword to 'Raw Data' Is an Oxymoron. In *'Raw Data' Is an Oxymoron,* ed. L. Gitelman, pp. 167–172. Cambridge, MA: MIT Press.

Bowker, G. C. and S. L. Star. 1999. *Sorting Things Out: Classification and Its Consequences*. London: MIT Press.

boyd, d. m., and K. Crawford. 2012. Critical Questions for Big Data. Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society* 15(5): 662–79.

boyd, d. m., and N. B. Ellison. 2008. "Social Network Sites: Definition, History, and Scholarship." *Journal of Computer-Mediated Communication* 13(1): 210–30.

Epstein, S. 2008. Patient Groups and Health Movements. In *The Handbook of Science and Technology Studies,* eds. E. J. Hackett, O. Amsterdamska, M. Kynch, and J. Wajcman, pp. .499-539. London: MIT Press.

Faraj, S., S. L. Jarvenpaa, and A. Majchrzak. 2011. Knowledge Collaboration in Online Communities. *Organization Science* 22(5): 1224–39.

Gerlitz, C. and A. Helmond. 2013. The Like economy: Social buttons and the data-intensive web. *New Media & Society* (OnlineFirst).

Hanseth, O., E. Monteiro, and M. Hatling. 1996. Developing Information Infrastructure: The Tension Between Standardization and Flexibility. *Science, Technology & Human Values* 21(4): 407–426.

Hayek, F. A. 1945. The Use of Knowledge in Society. *The American Economic Review* 35(4): 519–530.

Heidegger, M. 1962. *Being and time*. Oxford: Blackwell.

Howe, J. 2008. *Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business*. New York: Crown Business.

Jacob, E. K. 2004. Classification and Categorization: A Difference that Makes a Difference. *Library Trends* 52(3): 515–540.

Kallinikos, J. 1993. Identity, Recursiveness and Change: Semiotics and beyond. In *Tracing the Semiotic Boundaries of Politics*, ed. P. Ahonen, 257–78. Berlin: Mouton de Gruyter.

Kallinikos, J. 1999. Computer-based technology and the constitution of work: a study on the cognitive foundations of work. *Accounting, Management and Information Technologies* 9(4): 261–291.

Kallinikos, J. 2006. *The Consequences of Information: Institutional Implications of Technological Change*. Northampton, MA: Edward Elgar Publishing.

Kallinikos, J. and N. Tempini. 2011. Post-material Meditations: On Data Tokens, Knowledge and Behaviour. Paper presented at the 27th EGOS Colloquium - European Group of Organizational Studies, Gothenburg, Sweden, July 6-9.

Kallinikos, J. And N. Tempini (Forthcoming). Patient Data as Medical Facts: Social Media Practices as a Foundation for Medical Knowledge Creation. *Information Systems Research.*

Latour, B. 2000. When Things Strike Back: A Possible Contribution of 'Science Studies' to the Social Sciences. *British Journal of Sociology* 51 (1): 107–23..

Leonelli, S. 2012. Classificatory Theory in Data-Intensive Science: The Case of Open Biomedical Ontologies. *International Studies in the Philosophy of Science* 26 (1): 47–65.

Löwy, I. 2011. Labelled Bodies: Classification of Diseases and the Medical Way of Knowing. *History of Science* 49: 299–315.

Majchrzak, A., S. Faraj, G. C. Kane, and B. Azad. 2013. The Contradictory Influence of Social Media Affordances on Online Communal Knowledge Sharing. *Journal of Computer-Mediated Communication* 19(1): 38–55.

Mayer-Schönberger, V., and K. Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.

Mathiassen, L., and C. Sorensen. 2008. Towards a Theory of Organizational Information Services. *Journal of Information Technology* 23 (4): 313–29.

Millerand, F. and G. Bowker. 2009. Metadata Standards: Trajectories and Enactment in the Life of an Ontology. In *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life,* eds. M. Lampland and S. L. Star, pp. 149–165. Ithaca, NY: Cornell University Press.

Mingers, J. 2004. Real-izing information systems: critical realism as an underpinning philosophy for information systems. *Information and Organization* 14(2): 87–103.

Rabeharisoa, V., T. Moreira, and M. Akrich. 2013. Evidence-Based Activism: Patients' Organisations, Users' and Activist's Groups in Knowledge Society (CSI Working Papers Series. Working Paper 033). Paris, France: Centre de Sociologie de l'Innovation, Mines ParisTech.

RARE. 2014. RARE Facts and Statistics. *The Global Genes Project*. Retrieved from http://globalgenes.org/rarefacts/ (accessed Novermber 23, 2014)

Ribes, D., and G. C. Bowker. 2009. Between Meaning and Machine: Learning to Represent the Knowledge of Communities. *Information and Organization* 19 (4): 199–217.

Ribes, D. and S. J. Jackson. 2013. Data Bite Man: The Work of Sustaining a Long-Term Study. In *'Raw Data' Is an Oxymoron*, ed. L. Gitelman, pp. 147–166. Cambridge, MA: MIT Press.

Rose, N. 1999. *Powers of Freedom: Reframing political thought*. Cambridge: Cambridge University Press.

Rose, N. 2007. *The Politics of Life Itself. Biomedicine, Power, and Subjectivity in the Twenty-First Century*. Oxford: Princeton University Press.

Runde, J. 1998. Assessing causal economic explanations. *Oxford Economic Papers* 50(2): 151–172.

Sayer, A. 2000. *Realism and Social Science*. London: Sage.

Scheuermann, R. H., W. Ceusters, and B. Smith. 2009. Toward an Ontological Treatment of Disease and Diagnosis. In *Summit on Translational Bioinformatics 2009*, pp. 116–120. Available at: http://www.ncbi.nlm.nih.gov.proxyiub.uits.iu.edu/pmc/articles/PMC3041577/ (accessed November 23, 2014).

Shirky, C. 2008. *Here Comes Everybody: The Power of Organizing Without Organizations,* Vol. 86. London: Penguin Press.

Shirky, C. 2010. *Cognitive Surplus*, *Culture*. London: Penguin Press.

Star, S. L. 1983. Simplification in Scientific Work: An Example from Neuroscience Research. *Social Studies of Science* 13(2): 205–228.

Star, S. L. 1986. Triangulating Clinical and Basic Research: British Localizationists, 1870-1906. *History of Science* 24(1): 29–48.

Star, S. L. and M. Lampland. 2009. Reckoning With Standards. In *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*, eds. M. Lampland and S. L. Star, pp. 3–24. Ithaca, NY: Cornell University Press.

Timmermans, S. and M. Berg. 2003. *The Gold Standard. The Challenge of Evidence-Based Medicine and Standardization in Health Care*. Philadelphia: Temple University Press.

Timmermans, S., G. Bowker, and S. Leigh Star. 1998. The Architecture of Difference: Visibility, Control, and Comparability in Building a Nursing Interventions Classification. In *Differences in Medicine: Unraveling Practices, Techniques, and Bodies,* eds. M. Berg and A. Mol, pp. 202–225. London: Duke University Press.

Topol, E. 2012. *The Creative Destruction of Medicine*. New York: Basic Books, p. 303.

Treem, J. W., and P. M. Leonardi. 2012. Social Media Use in Organizations: Exploring the Affordances of Visibility, Editability, Persistence, and Association. *Communication Yearbook* 36: 143–89.

Turner, M. R., P. Wicks, C. A. Brownstein, M. P. Massagli, M. Toronjo, K. Talbot, and A. Al-Chalabi. 2011. Concordance between site of onset and limb dominance in amyotrophic lateral sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry* 82(8): 853–854.

van Dijck, J. 2013. *The Culture of Connectivity: A Critical History of Social Media*. New York: Oxford University Press.

Wicks, P. 2007. Excessive yawning is common in the bulbar-onset form of ALS. *Acta Psychiatrica Scandinavica* 116(1): 76–76.

Wicks, P. and G. J. MacPhee. 2009. Pathological gambling amongst Parkinson's disease and ALS patients in an online community (PatientsLikeMe. com). *Movement Disorders* 24(7): 1085–1088.

Wicks, P., T. E. Vaughan, M. P. Massagli, and J. Heywood. 2011. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nature Biotechnology* 29: 411-414. Available at: http://www.nature.com/doifinder/10.1038/nbt.1837 (accessed November 23, 2014)

Wikipedia. 2014. Evidence of Absence. *Wikipedia*. Available at: http://en.wikipedia.org/w/index.php?title=Evidence_of_absence&oldid=606482435 (accessed May 5, 2014)

Williams, T. D. 2013. Procrustean Marxism and Subjective Rigor: Early Modern Arithmetic and Its Readers. In *'Raw Data' Is an Oxymoron,* ed. L. Gitelman, pp. 41–60. Cambridge, MA: MIT Press.

Wittgenstein, L. 1953. *Philosophical Investigations*, (G. E. M. Anscombe, Tran.), Oxford: Blackwell.

Wynn, D. and C. K. Williams. 2012. Principles for Conducting Critical Realist Case Study Research in Information Systems. *MIS Quarterly* 36(3): 787–810.

Yin, R. K. 2009. *Case Study Research. Design and Methods. Fourth Edition*. London: Sage.

Zuboff, S. 1988. *In the Age of the Smart Machine: The Future of Work and Power*. New York: Basic Books.