**LONDON SCHOOL OF ECONOMICS**
**AND POLITICAL SCIENCE**

# GOVERNING SOCIAL MEDIA

Organising information production and sociality through open, distributed and data-based systems

Niccolò Tempini

A thesis submitted to the Department of Management of the London School of Economics and Political Science for the degree of Doctor of Philosophy, London, October 2014

*'I think what it boils down to, really, is that I hate reality.*
*And you know, unfortunately,*
*it's the only place where we can get a good steak dinner'*
Woody Allen

# Governing Social Media

Organising information production and sociality through open, distributed and data-based systems


## Abstract

This thesis explores the management of social media networks through a specific interpretive lens. It views social media as networks organised for information production and managed through the development of complex data structures and underpinning technological solutions. The development of social media networks – chiefly characterised by the open and distributed participation of many diverse individuals through the intermediation of specific technological solutions – seems to give shape to new organisational forms and data management practices, impacting in many domains. Despite vivid interest in these participatory organisational forms, we do not fully understand how social media technology is leveraged to organise member communities, standardising processes and structuring interaction. In this research I build on the case of *PatientsLikeMe*, a prominent and innovative social media network constructing medical scientific knowledge through the data-based contributions of its open and distributed member base. By drawing on the findings of an intensive, participatory case research, the thesis makes a contribution on several levels.

The thesis demonstrates that the management of social media networks is characterised by the need to achieve steady, reliable and comprehensive production of information and associated data collection by means of complex data architectures and user reporting. I illustrate these conditions by highlighting the challenges that characterise the development of a system able to engage productively with the member base and by describing the mechanisms and techniques through which the organisation seeks to address them. Data and data structures figure prominently throughout the research as organisational devices of critical importance for the management of social media networks.

The thesis also indicates and comments on the implications of these innovative modes of organising knowledge production. It finds that social media support considerable innovation in the arrangements by which scientific knowledge can be produced, with a consistent inclusion of once marginalised actors in data management practices, and elaborates on effects on the relationship with research institutions and professions. At the same time, the thesis shows that social media technology, because of the challenges and strategies associated with information production, ambiguously supports the project of a wider inclusion that it seems to afford at first sight. Finally, the thesis claims that developing social media gives rise to specific techniques of construction and governance of the social, and the associated kinds of sociality where socialisation, computation and the production of knowledge objects are inextricably enmeshed.

## Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).



I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 68575 words.

**Statement of conjoint work**

I confirm that the first paper was jointly co-authored with Aleksi Aaltonen (LSE) and I contributed 40% of this written work.

I confirm that the third paper was jointly co-authored with Jannis Kallinikos (LSE) and I contributed 60% of this written work.

**Statement of use of third party for editorial help**

I confirm that my thesis was copy-edited throughout, by either Emily Stapleton or Jonathan Sutcliffe, for conventions of language, spelling and grammar.

## Acknowledgements

This work has been made possible through the help and support of numerous friends and colleagues, more than I will be able to thank here. First of all I would like to thank my supervisor Jannis Kallinikos who has supported the project throughout from its early phases. I have enjoyed his precious guidance, when he suggested directions and warned from possible dead ends, and precise thinking, with his well-known attention to detail. Also, his work has inspired this work greatly. I have learned greatly from him, and will keep learning through the memory of the experience that I will carry with me. Thank you for all this, Jannis.

At the LSE Information Systems and Innovation Group I have also had the luck of two other friends and mentors, Carsten Sorensen and Antonio Cordella. Both of them have supported my work, reading and commenting drafts and encouraging me. I felt honoured when they took me on board to teach with them on multiple courses. Most importantly, they are very good friends of mine. Thank you so much, Carsten and Antonio.

A very special thanks goes to Aleksi Aaltonen. I have often joked that he has been an "academic big brother" to me. From the late 2009 days, when I was an MSc student in his class, to the early 2014 co-authored article (passing through an initiation to rock climbing, and others), we have done lots of things together. Thanks Aleksi for your kind friendship, I hope there will be many other occasions. Martha Poon has been another very special friend, with whom I spent countless nights of intense and brain-wrecking discussions. Thank you Martha, you will become an honorary Italian soon.

Friends and colleagues at the Department of Social Science, Health and Medicine, King's College of London, have also helped greatly in making this thesis what it is. Through the affiliation, I have had the possibility to present my work several times, receive comments and suggestions on drafts and get to know many other researchers through the Citizen Participation in Science and Medicine network. This work is especially indebted to Nikolas Rose and Barbara Prainsack – their help has been crucial. But also I would like to thank Claire Marris, Tamar Sharon, Ilana Löwy, Tara Mahfoud, Des Fitzgerald, Gísli Pálsson, Thorvald Sirnes and Silvia Camporesi, among others.

At the ISIG and LSE I have found many friends. I have taught with Will Venters, Steve Smithson, Maha Shaikh and Spiros Samonas. Thanks all, it was great fun! A special place for DoM *paisà* Chiara Benassi and Enrico Rossi. Then my thanks go to my fellow Information Systems Research Forum coordinators, Florian Allwein and Silvia Masiero, but I wave also to Cristina Alaimo, Nuno Oliveira, Antti Lyyra, Kari Koskinen, Carla Bonina, Rohollah Honarvar, Boyi Li and Helene Lambrix – among others! Thanks to fellow TIGAIRians Attila Marton and Jose Carlos Mariategui. Thank you also very much to the DoM staff, especially Leo Beattie, Fran White and Imran Iqbal. I bothered you often!

This work has also had the luck of receiving feedback from many other friends of the wider Information Systems and Management community. Michael Barzelay, Ola Henfridsson, Sue Newell, Jochen Runde and Paul Leonardi, thanks for our discussions. Also thanks to my ICIS 2013 Doctoral Consortium fellows Mohammad Moeini, Koteswara Ivaturi, Brian Dunn, Ina Sebastian, Inchan Kim, John Qi Dong, and Liliana Lopez.

Thank you to all the people at *PatientsLikeMe* with whom I have worked but also spent nice spare time with! Thank you also to Sabina Leonelli – seeing a new project ahead was strong motivation for the final push.

There are friends and companions to whom I am and will be indebted as a human being. They are always with me no matter the distance. Natalia, Alan, Nuria, Filippo and Marta, Stefano, Marco, Nathalie, Andrea, Michele, Jean Daniel, Kalman, Mattia, Lorenzo, are only a few of the names I carry in my heart and am thinking of in this moment. And it now feels like this thesis has been taking all of my life during the four years of work. Actually, there has been more to my life than this in such time-lapse, but nonetheless an intense thought immediately goes to companions that life has had no more – Nella, Stefania, Stefano, ciao.

This thesis is dedicated to my parents, Miriam and Luigi. It is because of their unconditional love and support that I have been free to roam and make the mistakes that could grow me. Then write a PhD thesis ;-) I love you both.

## Original Papers

The thesis includes the following papers, listed in chronological order (both of writing and presentation):

1. Aaltonen, Aleksi and Tempini, Niccolò (2014). Everything counts in large amounts: a critical realist case study on data-based production. *Journal of Information Technology* 29(1): 97–110.
2. Tempini, Niccolò (in press). Governing *PatientsLikeMe*: information production in an open, distributed and data-based research network. *The Information Society.*
3. Kallinikos, Jannis and Tempini, Niccolò (2014). Patient Data as Medical Facts: Social Media Practices as a Foundation for Medical Knowledge Creation. *Information Systems Research* 25(4):817-833*.*
4. Tempini, Niccolò. Till Data Do Us Part. Sociality and the proliferation of medical objects in social media-based discovery. *Manuscript under review for international journal: London School of Economics and Political Science.*

All of the work has been carried out following my registration in the PhD in Information Systems program in the Information Systems and Innovation Group, Department of Management, LSE.

# Table of Contents

# Introduction

This thesis explores the management of social media through a specific interpretive lens. It views social media as networks that are organised for the production of information and managed through the development of complex data structures and underpinning technological solutions. Data are what a social media network routinely generates as users use the technology in their daily interactions, and it is from the analysis of such data that social media organisations can produce information about users and their life contexts. Data structures (models, classifications, categories) are the fundamental communication intermediaries connecting diverse and distributed contexts. They are the cognitive grids through which users are understood and analysed, fundamental tools that allow creating value and exchanging information through the networks, and as such they are at the centre of this study of social media.

Through this perspective, the thesis explains how and why social media organisations control and coordinate the activity of the network, as they try to elicit select user behaviour for the accomplishment of specific data generation and collection tasks. Also, the focus on data structures and technological intermediation explains how social media organisations are able to operationalize organisation forms for data generation and collection that are characterised by openness (broadly open to content from anyone and about anything), distribution (connecting with individuals from any context), and data-basedness (intermediation of objects and events of the world through translation in various forms of data).

There is little doubt that the unprecedented capacity of social media to dynamically attract open publics (van Dijck, 2013), made of diverse people distributed across the globe, and involve them with one another or with a shared goal has aroused public opinion, albeit often the focus of the fascination has been more concerned with fluid, creative and bottom-up interaction dynamics that these networks afford (boyd and Ellison, 2008; Faraj *et al.*, 2011; Kelly, 1996; Shirky, 2008, 2010). Not enough attention has been paid to the precise mechanisms and processes by which social media organisations govern their networks. Indeed, we understand that social media are engineered environments fostering particular forms of online sociality (van Dijck, 2013), but we need more to understand the specific forms and processes of social media management and development. We need thorough explanations that show precisely how social media are leveraged to organise member communities, standardising processes and structuring interaction. We need also to investigate their eventual or potential relationships with traditional data management practices, institutions and cultural forms.

So, what are the features of social media that make them unique? To sketch some essential elements in introduction, we must observe that these networks have consistently afforded the gathering and coordination of open and distributed user bases of unprecedented scale and across the whole social fabric (boyd and Ellison, 2008; Faraj *et al.*, 2011; Shirky, 2010; van Dijck, 2013). They have empowered users with new (mass) communication media. Outcomes have often been remarkable accomplishments, that could not have been achieved without such technological underpinnings. There are many examples in several different domains from gigantic endeavours of mass collaboration (e.g. Wikipedia, Ushahidi) to the unprecedented granularity of knowledge

peer production (Aaltonen and Kallinikos, 2013; Benkler, 2007; Shirky, 2010). The role of social media in underpinning approaches to production that have challenged established configurations of organisational boundaries and actors cannot be ignored. Also a prosaic example like Facebook, perhaps the epitome of arguments of trivialisation of personal interaction (Borgmann, 1999), has represented for its users, among other things, a near-irresistible tool for sharing and constructing memories with friends and loved ones. Social media have afforded a new range of unprecedented social interactions at the same time as productive outcomes (Treem and Leonardi, 2012). They are underpinning, as I illustrate in the thesis, a new kind of sociality, where computation of data generated through complex infrastructures of power shapes the connections and interactions that users of social media can entertain with others (Kallinikos and Tempini, 2014; Tempini, in press).

In this perspective, it seems reasonable to wonder what organisation forms must underpin these outcomes. How do these developments occur? A popular myth is that these most admirable phenomena have been spontaneous and unexpected. At the same time, the mainstream success of social media has made all of us aware that major organisations own the most prominent social media networks and try to reap financial benefits out of them (van Dijck, 2013). Quite obviously, the main focus of the organisations' efforts is the continuous and incremental development of the social media technology. Taking over the control of projects that on most occasions started informally (for instance as student projects, and hobbies), social media organisations put the platform at the centre of the business. Several different business models have been tried in order to make social media platforms sustainable (van Dijck, 2013), but we know that critical mass, user lock-in and monetisation have been difficult obstacles

to surmount, and failures in the social media space have been numerous and have corroborated the famous Silicon Valley principle of 'fail fast, fail early, fail often'.

An important common denominator of the social media space has been, throughout its brief history, that most if not all social media start ups execute business models centred on the sale of services that are fuelled by information generated through the social media network (Aaltonen and Tempini, 2014; boyd and Ellison, 2008; van Dijck, 2013). In social media, the network's capacity to generate information on a user base (who the users are; what they do, believe or like) is tightly interconnected with its capacity to engage and control it (van Dijck, 2013). Part of the explanation of the social media phenomena and their spectacular outcomes must therefore be connected to a two-fold relationship with its user-base. On one hand the organisation needs to give its users what they need in order to have fulfilling experiences and social interaction. On the other hand, the organisation needs to be able to collect the kind of data that is needed for information production needs. In social media but also, as I show in the first paper of the thesis, other ventures where business models depend on the production of information routinely generated by a network infrastructure, digital data are the raw matter that stands at the centre of work processes (Aaltonen and Tempini, 2014). Needless to say, great corporate empires have been developed essentially on this basis.

Given this purview, one should ask what resources and strategies does a social media organisation have at its disposal to relate to its users and govern the network towards desired levels of productivity? In a completely intermediated social environment, the governing power of technology is the first and paramount element of

examination (Kallinikos, 2011). It appears that technology is the main tool, or the umbilical cord, that the organisation can use to grow and control the social media network user base. No in-depth analysis of social media can take place without a serious account of technology and its role in this respect. As the Information Systems discipline has demonstrated, this is not an easy topic. The possibility of exploiting technology to govern the behaviour of users is an old theme of instrumental rationality and technological determinism and, notoriously, it has proven false on numerous occasions (Ciborra, 2002; Markus, 1983; Orlikowski, 1996). At the same time, it has been demonstrated that something is at work when standard routines and work patterns are embedded in large packages that are then forcedly implemented (Kallinikos, 2006, 2011; Pollock and Williams, 2009). The state of the debate on the relationship between technology and social contexts can perhaps be broadly summarised by saying that while technology has been demonstrated to be unable to determine social (e.g. to get users to do what is intended), it has also been ascertained that technology can impose constraints (e.g. it can be quite effective in impeding users from doing what technology is designed to forbid). The weakness of technology in social determination is also due to not only dissociative behaviour by the end user – resistance, work arounds, amongst others – but also and importantly to the difficulty in devising reductions of the social world that successfully align across contexts (Bowker and Star, 1999; Ciborra, 2002; Hanseth *et al.*, 1996; Jacob, 2004; Mathiassen and Stage, 1990). The importance of this factor has proven greater as technological systems become more distributed and involve a larger diversity of contexts and social actors. At the level of information infrastructures, developing technology for controlling users may not reach the desired results even in the face of collaborative end-users. Given these preliminary observations we can perhaps say, with a safe degree of approximation, that purposeful development

of technology is a potentially useful strategy for governing user bases, but one characterised by a high degree of contingency. In this respect, the findings of the thesis highlight new forms by which this theme is perpetuated. In respect to the technology underpinning social media, one more premise is in order. Social media are typically characterised by their reliance on web-based technologies and Internet networking. While discussing the very specific technical features of these technologies is not a concern of the thesis, it should suffice here to say that a new generation of web-based technology has allowed the building of complex data architectures, and that data architectures are a central organisational device to connect across various social contexts and engage distributed users so broadly (Millerand and Bowker, 2009; Morville, 2005; Ribes and Bowker, 2009; Smith, 2008). It follows that in understanding the role of technology in shaping and organising social media networks, particular attention must be paid to data structures and their complexities. The argument for this stance is properly articulated in the papers that make up the rest of the thesis.

I conclude this section by reformulating in a few questions what emerges in light of these initial considerations. The perspective from which I started and conducted this research, outcome of both my wonder and reflection, was to understand the tripartite relationship between the social media organisation, the technology and the users. It can be further broken down in these questions:

- *How does a social media organisation develop technology in order to govern the platform's user base?*
- *How does social media technology support the management of the network?*
- *How does a social media organisation manage its information production process?*

- *How are social media networks shaping sociality and the users' life context?*

These questions are answered to throughout the thesis. In what follows, I recapitulate the most important topics that traverse the portfolio of papers that constitutes the body of the work.

## Topics

In the thesis I argue that social media technologies cannot be fully understood in their social significance unless we adopt a particular perspective on organising through social media – as geared towards *information production*. I apply this perspective throughout the thesis. If we do not look at information as the *cognitive currency* (Kallinikos, 2011) that the social actors involved exchange through their interactions, we cannot fully explain how a social media organisation reaches a desired, productive equilibrium with the underlying user base and how it manages and develops the platform in order to maintain this equilibrium.

Social media are underpinned by technological infrastructures that allow the generation and transaction of information (between users; with the organisation) through digital data. Needless to say, sociality in an open, distributed and data-based system is based on information transactions. Data are generated and exchanged as users interact with each other, but also as the organisation interacts with or profiles the users. Information and the data it stems from are the raw matter (Aaltonen and Tempini, 2014) that allows the organisations to experiment with different business models – be they based on advertising, premium features, or sale of data for market or scientific research. Notoriously, in the most prominent mainstream social media

networks, organisations extract information from the behavioural traces and interactions that are recorded by the platform in the form of digital data, and sell either direct access to this information or derivative services (e.g. advertisement services networks). It is thus of paramount importance to answer questions such as *how do social media organisations go about producing information, and grow the social media platform?* If we look at the mechanisms of information production, I argue in the thesis, we gain a powerful viewpoint to understand how the organisation changes and develops the technology in an attempt to govern the network and its user base towards productive (information-productive) behaviour.

In social science, to respond to questions about the relationship between technology and society, or between technical and human elements, often requires connecting multiple and apparently separate fields of inquiry. Social media open up the issue of organising information production beyond the boundaries of formal organisations, reaching out to recondite niches of the social. It becomes evident then that the research topic of the management of social media platforms is of paramount importance as it is at once a topic of technology and information production and a broader topic of technology and society. It is a goal of the four papers to identify some of the most important issues that are connected to these organisational forms, which reach beyond the organisation through open, distributed and data-based infrastructures and put digital data objects at the confluence of the efforts and interactions of many social actors.

In the following sections, I briefly summarise the main thematic strands that interconnect the papers to each other under the horizon of 'social media as systems for

information production'. Social media are studied here from an organisational perspective – I look primarily at what the sponsoring organisation does or can do with such infrastructures – a different view from the perhaps more popular one where the preoccupation has often been to see what users can do with the social media technology (boyd and Ellison, 2008; Faraj *et al.*, 2011; Treem and Leonardi, 2012; van Dijck, 2013). By drawing on the findings of two intensive, participatory case research studies (one pilot and one main case), the thesis makes a contribution on four main levels.

***Social media as open, distributed and data-based systems for information production***

First, the thesis illustrates how the management of social media networks is characterised by the need to achieve steady, reliable and comprehensive production of information and associated data collection by means of complex data architectures and user reporting. I illustrate these conditions first by highlighting the challenges that characterise the development of a system aimed at productively engaging the member base and second, by describing the mechanisms and techniques through which social media organisations seek to address them.

In this line of research I construct through my empirical work a ground for linking and comparing with the case of another organisation that exploits similarly open, distributed and data-based architecture that does not however belong to social media – the network does not strictly resemble social media, lacking their community structure, social functions and affordances. However, like social media companies, such an organisation fully revolves around the collection of informative data, and the

sustainability of its very different business models invariably depends on the function of discovery and reporting about distant and unknown contexts and actors. In connection to this, and most importantly for the thesis, this other data-centric organisation does not manage and develop data architectures of comparable complexity to those powering social media networks, nor requires or aims to have comparable levels of understanding and knowledge of the network users and their life contexts.

In the first paper, I explored the pilot case of an organisation exploiting the data routinely generated by a mobile network infrastructure with the aim of extracting information about the network's user base. I started to develop a conceptual framework on data-based information production that I have continued to develop in the following three papers, which are about the case of a social media organisation. The central concern, in social media organisations as well as other organisations that maintain large infrastructures to connect with distributed and largely unknown users and their contexts, is in gathering, maintaining and governing the network to elicit select behaviour. The result of the investigation in this topic is the creation of an original concept to capture the organising logic of open, distributed and data-based approaches to information production, that of *information cultivation*. In the other main case, based on social media platform *PatientsLikeMe*, I elaborate on this perspective and show how the production of information about the network user base involves explorative efforts of objectification of social objects and actors that remain beyond direct access – what I define, with import from arithmetics, as *social denomination*.

Among other essential features, the information production perspective is founded upon Bateson's (1972) event-theory of information – information as a context-

dependent event yet associated to context-independent semantic features of the chosen system of signification. In this respect, the eventual capacity of data to be equally informative notwithstanding different contexts and different observers makes information context-dependence a seemingly paradoxical property (Kallinikos, 2006), and my work tries to evolve under the horizon of this problematic, especially in the second paper. This first broader topic on social media organising underpins all papers of the thesis, as also the last two papers on social media build arguments that depend on the perspective on social media development and organising that I have mostly laid out in the first two.

***Data pools and data structures as management organisational devices***

At the centre of the explanation of the efforts of governing a user base for the production of information, the paramount role of data and data architectures in social media systems powerfully emerges as the second paramount topic of the thesis. Data structures are here fundamental intermediaries of social interaction, shaping the mutual understanding of social actors as the organisation tries to discover and construct knowledge about the users, and the users try to get to know and learn from each other. The resulting pools of social data as captured by the social media infrastructure are clearly the resource that makes a whole social media-based business model possible. Due to data's intrinsic context-dependence, the potential information that the data contains may well not be realised, and consequently organisational efforts revolve around the outcomes of data sense-making.

The role of data structures as gateway technologies, connecting multiple actors to distributed contexts and effecting an imposition of sameness between different

objects or individuals – making things countable – has been anticipated in social science (Bowker and Star, 1999; Kallinikos, 1995; Ribes and Bowker, 2009). The literature on data and classification systems in distributed science has indicated and fleshed out this role when scientific enterprise – explorative by definition – is conducted which connects distributed scientists and contexts to each other. In this respect, the original contribution of my work is not only to show that data structures are real structures in their own right, shaping organisation in specific ways. It is also to show how in social media (arrangements open to an undefined variety of phenomena to be reported by an undefined user base) the constant tinkering and manipulation with data structures happens at unprecedented dynamism and for the aims of governing social representation and interaction. The organisation - in a fast-paced feedback loop where the assessment of the collected data's potential to inform shapes the further modifications carried out on the data structures - actively governs the community it sponsors and fosters.

This topic underpins the four papers. While the first paper, centring on a pilot study of a non-social media organisation, fleshes out the organisational mechanisms of data-based information production, it is with the contrast offered in the following three papers on the social media case of *PatientsLikeMe* that it is shown that data structures of different complexity and models differently shape the productive relationship with the network's user base. In particular, the power of data structures in organising data collection is assessed and made sense of through the aggregation of vast data pools. In this purview the thesis shows the precise dynamics and operations through which big pools of data are made sense of and become a most important link in the chain of development iterations evolving a social media platform and shaping its sociality. The

thesis puts the management of data, data structures and data aggregates at the centre of the study, avoiding a limiting focus of an analysis concerned only with the status of the static knowledge representations embedded in data structures.

***Developing productive and digital sociality***

The third topic prominently emerging in the thesis and overarching the three empirical papers on the social media case shows how developing data structures for information-productive data collection involve the construction of the online sociality unfolding through these large end-user systems. First, by showing the processes of the development of data structures in social media, I reveal how categories and classifications shape representations of the social within the community itself, changing the meaning of objects and the representation of member individuals. Second, I show how the system is engineered to lead to select data collection behaviours through the analysis and computation of the available user data with the aim of offering differentiated opportunities for consumption of information and associated social interaction. Importantly, I show that the system thus aims at creating a loop where data sharing behaviours follow one after the other through the fuelling and shaping of sociality.

Through the concept of social denomination, as constructed in particular in the second and fourth paper, I describe a technique by which categories of social objects or subjects are defined, dynamically and through data constructs. As such, I show how the organisation managing the social media network is in a position to continuously reshape the cognitive foundations upon which the system works and aggregates individuals and their data. The technique has two main consequences. First, to establish

the cognitive common denominators that afford the computational operations that lead to data aggregation, a fundamental step for assessing the status of the data collection that the social media infrastructure generates for the owner organisation. Second, to guide or entice individuals to information-producing behaviour through continuously updated representations of the self and of entities from the patient's life context.

Through the interrelated concept of computed sociality, instead, I have developed this theme one step further, to show how the system uses social data – as shaped by the data structures through which the data are collected – to continuously draw and re-draw, through computation and representation techniques, connections and interaction opportunities with other individuals. This is identified as a key process at the heart of the engineering of online sociality (van Dijck, 2013), and the third paper provides the necessary empirical detail to demonstrate the argument. The concepts of social denomination and computed sociality interweave, in that social denomination describes the operations through which the computed sociality is shaped and constructed.

The empirical findings, showing developers facing complex data modelling dilemmas, demonstrate unambiguously how these issues rest at the heart of the socio-technical shaping of sociality that social media systems engender. Data and information are not only organising information production, but also and at the same time organise social interaction and representation.

***Making countable people and objects count***

The fourth and last topic in which the thesis indicates and comments on the larger implications of innovative modes of organising knowledge production through social media, such as the one analysed in the *PatientsLikeMe* empirical case. I find, in the third paper, that social media afford considerable innovation in the configurations by which expert (scientific) knowledge can be produced. I point at a consistent inclusion of once marginalised "lay" actors in data management practice and the eventual reshaping of relationships with and between incumbent institutions and professions. I demonstrate how the inclusion of the "lay" user is afforded by a very precise and continually refined architecture of data collection work. Also corroborated by the empirical evidence presented in the second and the fourth paper, the thesis shows technology to have a fundamental role in breaking down and framing data collection tasks into units that are manageable by the "lay" without continuous "expert" supervision. The entire social media-based business model, and the innovation it represents, rests on specific technological solutions that make it possible for a new data collection architecture to be executed and break away from the expensive arrangements of traditional research approaches. The thesis, however, not only establishes a strong link between technology and a specific organisational outcome, but also points at consequences for work boundaries and reconfiguration of roles (Abbott, 1988; Gieryn, 1983; Jonsson *et al.*, 2009).

In addition, in the fourth paper I show how social media technology, because of the challenges and strategies associated with information production which have been illustrated, only ambiguously supports the project of a wider inclusion of marginalised actors in technology development and scientific knowledge-making. Through the

empirical findings, I demonstrate that the continuous manipulation of categories and classifications powering a social media infrastructure often has destabilizing consequences for social interaction and representation. Showing that objects and subjects participating in social media communities are provided a foundation precisely through the cognitive devices of categories and classifications, the arguments follows demonstrating that continuously shifting these foundations, as the organisation's knowledge of the community and its members evolves – by social denomination – means to shift not only the basis upon which individuals and entities are *countable* but also how individuals are *made to count.* This is an important issue, marking an important difference between scientific research through social media and traditional distributed science efforts, traditionally involving many subjects and contexts but usually of known and finite number. The complexity in representing the social world in a social media network derives largely from its necessarily open character. The system needs to serve and adapt to contexts, phenomena and participants of essentially unknown diversity.

## Research Approach and Methodology

To address and connect to these multiple areas of inquiry I draw from work originating in different disciplines, including, in addition to information systems, sociology, information science and philosophy. I argue with others (e.g. Hanseth, 1996; Sayer, 2000) that explorations of new social phenomena often require interdisciplinarity – a potentially risky position. The information systems discipline itself, on the forefront of the understanding of some of the most prominent developments of our time, has often characterised itself as an essentially interdisciplinary field. This, one might say, is not surprising for a discipline that has

made of modernity and some of its most prominent consequences its elective turf (Feenberg, 2010; Kallinikos, 2006). Still, there are liabilities involved in interdisciplinary scholarship and it is commonplace to consider it difficult to deliver high-standard research. In addition, interdisciplinary work can be exposed to multiple fronts of critique and cannot exhaustively master each of the fields it draws upon. The most rational strategy for controlling for these liabilities is to carefully open up the new phenomena and associated problematics by selectively choosing the theoretical and empirical resources. These must be based on one's own position as a researcher, his or her competence and the different priorities one would attribute to a broad set of issues. Complete exhaustion of the issues regarding social media in society is clearly not possible.

To understand the tripartite relationship between an organisation, the technology and its users, I have specifically chosen to focus on the production (and consequent exchange) of information as this is the most important perspective to explain the connection between the parties involved. Indeed, in social media-based arrangements where one organisation depends on the value that it can extract from the activity of the network members, no monetary exchanges are usually involved. Quite literally, then, information is the currency (Kallinikos, 2011:72) that fuels the relationship between the parties involved. Information is the effectively transacted value passing hands between the components of a system.

The perspective that information production stands at the centre of data-based, network organisations – and the precise differentiation between data and information that this stance implies – requires to take a stance on a theory of information. In this

regard I build upon Bateson's theory of information, which sees information as an event that stems in context and not as an independent object or resource (as it is sometimes unfortunately portrayed by mainstream literature). Meaning depends on a connection between subjects and signifying objects, and information arises from the marking, through systems of signs, of differences between objects or states of the world. The theory affords the concept of information a particular depth, as a phenomenon that emerges depending both on context-dependent and context-independent features, and as such it is particularly suited to studying organisations where data are generated in one context and by the hand of one actor, and need to be made sense of – to become informative – in another context and by someone else.

Through an initial review, it emerged that the state of the field in respect to data, data pools and distributed networks (including social media) does not seem to provide a strong framework or set of structured research propositions with which to study social media from the perspective of an organisation developing the technology in order to govern a user base. IS literature has dealt with the topic of social media, but has mainly explored how social media technology affords a range of types of connections between social actors, or how the user base of online communities can prolifically generate emergent innovations. Meanwhile, scholarship on distributed science and categorisation systems, largely belonging to the field of science and technology studies, has not translated this body of knowledge over to social media phenomena. It is a contribution of the thesis to link the two bodies of literature by borrowing from the social study of classifications and categories. In this respect, I have borrowed a set perspectives and sensitivities in order to give shape to my case study design, and I contribute back to the literature with the specific findings of the cases; and with a new

combination of this body of scholarship with a critical realist framework centred on the Batesonian event-theory of information.

***Methodologies and Research Contexts***

From an evidence-making perspective my aim was to provide a process explanation of social phenomena that tightly connects what users do on the platform as they interact and socialise with the goals of the developing organisation, with a particular interest on data management practices and the intermediating role of data and data structures. It became necessary to equip myself with methods apt for the exploratory study of process. Case study research is particularly suited to this kind of exploratory investigation, as it allows the inclusion of a very wide and varied range of variables in the explanation of events and relationships between entities. The method allows the researcher to follow phenomena closely from within a rich context of observation, and keep the explanation open to the theorisation of new social mechanisms not prescribed or assumed by existing theories. For these characteristics, when it is combined with observational or ethnographic data collection the method is well suited to the construction of causal chains such as those that I have been constructing in the papers.

Based on these considerations, and in order to compensate for the lack of strong theoretical frameworks to direct the research with, I adopted Critical Realism as a meta-theory, a tightly-knit theory of ontology and epistemology that has a strength in constructing explanations from day to day events and operations. The view has been that providing myself with the strongest among "generic" research tools can enable robust theorisation of small and mid-range phenomena. While Critical Realism can be

combined with any kind of methodology, it works particularly well with observational, intensive case studies (Sayer, 2000; Wynn and Williams, 2012). In particular, retroductive theorising is a great resource for explanation building (Mingers, 2004) that is perhaps underemphasised or not properly highlighted by scholars working with other frameworks. It can support robust theorising through hypothesis of counterfactuals (Runde, 1998). In the case study in the first paper, I have argued for the methodological fit of a combination of Critical Realist meta-theory and observational case study research. I have stressed the role of retroductive theorising in particular in combination with analytical writing as a specific technique to aid the process. This whole research toolkit has then been consistently applied throughout the other three papers that complete the thesis.

*Pilot case*

The chosen methodology allowed me to prepare towards the case study on social media through a pilot case study. This was an important preliminary step because it allowed me to establish and test the investigative framework on information production. The other important value of the pilot case study is to demonstrate the methodology (including the application of analytical writing and retroductive process) that I have consistently applied throughout the rest of the thesis, as demonstrated in the methodology sections of each paper. The first draft of the resulting pilot case paper was submitted as the data collection for the social media case study was starting. The pilot case is at the centre of the first article of the four making up this thesis, while the other three articles report on the main case - social media.

The pilot case looks at information production in the context of an organisation that has not developed a social media network, but bears very important similarities to the social media organisation at the centre of the main case study. The case organisation is a mobile virtual network operator, which controls a mobile network infrastructure and tries, through it, to construct a media audience for the sale of advertising services. This organisation leverages digital media and a distributed infrastructure to run an innovative business model centred on the production of information from the data that is generated by the behaviour of a broad and dispersed user base. Similar to social media, the network's user base is largely undefined and practically out of reach of the organisation. In addition, the viability of this whole enterprise depends on the ability to extract information from the large amounts of data that the network produces.

The pilot study analyses its field site with the particular focus on data practices that is at the centre of this thesis and which I continue to develop in the following articles about the main case study. The article theorises mechanisms of information actualisation that explain how an organisation works through large pools of data to extract business-critical information. It demonstrates that data pools are not just heaps of data, instead they are real structures in their own right, with properties and powers of their own. They shape organisational work processes in specific ways. The article advances the understanding of many innovative organisational settings that are centred on new practices of data processing, such as social media organisations.

*Main case*

Important parallels can be drawn between the pilot mobile virtual network operator and social media organisations, and it is against this background that I have

organised and conducted the following main case study. The pilot case study constituted a benchmark that allowed specificities of social media organisations to emerge more easily. Following guidelines that I set out in the first paper (Aaltonen and Tempini, 2014), I focused on 1) events that are essential for organisational survival and 2) on the techniques of measurement of the public the network engages in. I approached the field with the aim of understanding the role of technological structures in supporting processes in the organisational setting that depended on distributed involvement of an undefined user base, with a specific interest in the management data structures and the related governing of the user base towards select behaviour. In a similar way to the media audience that the pilot case organisation constructs through a mobile network, the social media organisation's main concern is in gathering, maintaining and understanding about a user base. The main case organisation is a very particular one, producing scientific research with the information it generates through the social media network it maintains and the collaboration of patient users. The domain of medical research is one where data requirements are particularly specific and arrangements for scientific production are well-established.

It clearly emerged from the beginning of the study that a very important difference between the pilot and the main case stands in the complexity and forms of the data structures. The social media system embeds complex data structures in order to engage with the user base and select behaviour at a much higher granularity. The three *PatientsLikeMe* study papers show that new challenges emerge as social media organisations deal with much more complex data pools – made of structured data and interconnected categorisation and classification systems. The complexity of the data structures that the social media organisation manages were in close association with

the much more granular and comprehensive understanding of the network's users and their life contexts that the organisation was trying to construct. The data that the pilot organisation manages are instead less descriptive and unstructured. As such, it is suggested that the information production challenges identified in the main case study seem to be specific to the social media infrastructure. However, it is important to note that the findings from the pilot case supported and shaped the perspective through which the main case was conducted and consequently its findings. User base behaviour continued to silently and resiliently surprise the employees of both organisations, as they tried to form an understanding about the out of reach world through the intermediation of the data. Data do not determine the event of information but equally, information could not be realised (actualised – as in Aaltonen and Tempini, 2014) without data.

The paper portfolio as a whole provides a unique perspective on the management of social media networks centred on the perspective of information production and an empirical account of the development of data structures for the mobilisation of the information potential that a network of connected users might express. The portfolio develops this perspective through a set of tightly interconnected papers. The papers elaborate the topics in a progressive "escalation" from the explanation of small and idiosyncratic information production events up to the systemic and more broadly social and ontological consequences of social media-based arrangements for production. This thematic progression reflects the chronological progression of the work. The papers are now presented in the order they have been ideated, started and written. At the start of the conclusion chapter, I recapitulate them and summarize their individual contribution.

# Everything counts in large amounts: a critical realist case study on data-based production

**Aleksi Aaltonen and Niccolò Tempini, LSE**

**Abstract**

Contemporary digital ecosystems produce vast amounts of data every day. The data are often no more than microscopic log entries generated by the elements of an information infrastructure or system. While such records may represent a variety of things outside the system, their powers go beyond the capacity to carry semantic content. In this article, we harness critical realism to explain how such data comes to matter in specific business operations. We analyse the production of an advertising audience from data tokens extracted from a telecommunications network. The research is based on an intensive case study of a mobile network operator that tries to turn its subscribers into an advertising audience. We identify three mechanisms that shape data-based production and three properties that characterize the underlying pool of data. The findings advance the understanding of many organisational settings that are centred on data processing.

## Introduction

Prominent IS scholars have repeatedly complained about weak theoretical foundations for analysing the mutual constitution of technological systems, organisational arrangements and outputs (e.g. Lyytinen and Yoo, 2002; Orlikowski and Barley, 2001; Yoo, 2010). In order to cope with the problem, researchers continue to

import theories from other disciplines, whereas attempts to strengthen theory building capacity within IS are rarer (Baskerville and Myers, 2002; Benbasat and Zmud, 2003; Lee, 2010). In the spirit of the latter approach, this article demonstrates how critical realism (CR) helps to build a theoretical explanation of a specific, data-driven product innovation in commercial media. CR works as a metatheory[1] for our study. It is not concerned with specific empirical phenomena but is rather a theory of ontology and epistemology that guides the construction of theoretical explanations. Critical realism provides a robust, explicit framework for theorizing causal mechanisms that underpin a new kind of advertising audience.

The analysis revolves around a start-up telecommunications operator that has built a new form of commercial media by relaying advertisements to mobile phones as text and picture messages. The challenge for the company is that sending marketing messages to consumers does not yet constitute a viable medium for advertising. This is because advertisers are not willing to pay for advertising to an unknown audience (Ettema and Whitney, 1994; Napoli, 2003). Any aspiring media company must know its audience along relevant dimensions – otherwise it cannot sell media space to advertisers. This knowledge is typically based on a sophisticated technological capacity to monitor people's exposure to media content and advertisements. The opportunity for the company to construct an audience is grounded on its access to data from a

[1] By metatheory we refer to reasoning behind empirical research designs; a framework that provides the rationale and practical guidance on how the different aspects of research are brought together into a coherent argument. The term is largely synonymous with theoretical perspective (Crotty 1998), yet 'metatheory' communicates explicitly the idea of theory about research and distinguishes it, in our case, from substantive theorizing of technology in particular settings.

telecommunications network infrastructure. To understand the emergence of a new kind of advertising audience, we ask the question:

*What mechanisms allow the company to manufacture an advertising audience from the mobile network data?*

The idea of audience is a slippery concept that has no single accepted definition (Bratich 2005; Morley 2006). In this article, we understand an audience first and foremost as a product. The business of media companies is about creating, maintaining and selling audiences to advertisers. This is made possible by audience measurement arrangements, whose evolution has historically shaped media products, content and the whole industry (Bermejo, 2009; Carr 2008; Napoli 2003, p. 83). For this purpose, a mobile network infrastructure has the special feature of generating data tokens known as call detail records (CDRs); these capture network subscriber behaviour in a microscopic, standardized way across network elements. Yet, as we will show below, CDRs are meaningless in the context of organisational practices. No relevant pattern or insight emerges by looking at the raw data tokens. In order to have a product to sell for the advertisers, the company must turn the data into information about an audience.

The data tokens can be understood as non-material technological objects (Faulkner and Runde, 2009, 2010, 2013; Runde, Jones, Munir and Nikolychuk, 2009) or digital objects (Ekbia, 2009; Kallinikos, Aaltonen and Marton, 2013). The concept of object is central to critical realist theorizing and connects the study with recent discussions on materiality (Leonardi, Nardi and Kallinikos, 2012; Mutch, 2010; Orlikowski, 2007). We assume that the data tokens have syntactic properties that make a concrete impact on the audiencemaking operations. These properties neither derive

from the physical medium storing the data nor are simply representations of external reality. Indeed, we argue that the data are 'material' in the adjectival sense that they matter beyond their semantic content. Phenomena like those that we set out to investigate are the focus of what has also been called digital materiality (Yoo, Boland, Lyytinen and Majchrzak, 2012). Advertising audiences certainly have a lot to do with people using media content, but the variables that ultimately construct the audience product in the industry have always been influenced by technological measurement arrangements (Ettema and Whitney, 1994).

The article makes two contributions. First, we show how the critical realist framework supports the theorizing of causal mechanisms which are activated in the audiencemaking process (Bhaskar, 2008; Sayer, 2000). The term 'audiencemaking' is used throughout the article as shorthand for the construction of an audience as a product (Ettema and Whitney, 1994). Second, the properties of digital data and related causal mechanisms that emerge from them are not idiosyncratic to this case study. Given the relatively generic nature of data tokens such as log entries across different systems, our results can inform studies focusing on a wide variety of settings.

## Critical Realism

Critical realism (CR) is a philosophy of science that has a set of basic principles at its core (Archer, 1998; Bhaskar, 2008; Mingers, 2004; Sayer, 2000). The approach makes two fundamental assumptions with respect to the methodology of empirical research: first, the world exists independently from our knowledge; second, the world can be observed only partially. CR can be thus seen as drawing from the constructivist critique to earlier forms of realism, holding that both researchers and their informants

encounter the world through interpretation (Sismondo, 1993, p. 535). Importantly, however, CR also holds that those interpretations can carry traces of a reality that is independent of present actors. This allows CR to incorporate the idea that all knowledge is socially constructed and thus *transitive*, while scientific knowledge addresses *intransitive* structures of reality that do not depend on individual awareness of them and are independent from any given context. The difference between transitive knowledge and intransitive reality is central to CR and will be discussed below.

### *Transitive knowledge about intransitive reality*

There would be little point in CR if the intransitive reality simply mapped to natural phenomena while all artificial (Simon, 1996) were considered transitive. Quite the contrary, the intransitive reality is very much populated by the outcomes of human actions and interpretations. Let us call these relatively stable human-made entities 'social structures'.

All action depends on structures. Archer (1998, p. 197) points to Bhaskar, who "states unambiguously that 'social forms are a necessary condition for any intentional act, (and) that their pre-existence establishes their autonomy as possible objects of scientific investigation'". Social structures enable and shape actions, and as such they are important objects of scientific research. Entities such as a cultural convention, technological infrastructure or a law can have a structuring effect on action. Structures originate and are reproduced in human activities. Nevertheless, CR differs from popular IS approaches, such as structuration theory, actor-network theory and sociomaterial perspectives, in that it rejects the conflation of structure and action. An action cannot

draw upon a structure and simultaneously bring it into existence (Archer, 1982; Mutch, 2010).

The separation of action from structure is described in the transformational model of social activity (TMSA). The model describes how action draws upon, reproduces and changes structures in a temporal sequence (Faulkner and Runde, 2013; Runde, Jones, Munir and Nikolychuk, 2009). In our analysis, the focus is on the implications of an already existing structure (CDR data) on audiencemaking. We are interested in understanding mechanisms that emerge from the structure in a particular setting rather than in structural transformation. Consequently, we demarcate the case so that the construction of the CDR infrastructure is excluded from the analysis. This is also justified by the fact the infrastructure is taken as a given for all practical purposes at the research site. The CDRs are, in the language of TMSA, a structural condition for the company operations.

In the critical realist framework, CDR data objects, the instantiation of audiencemaking events and empirical observations map to different epistemological domains. The approach postulates an ontology in which the phenomena of scientific interest are structured beyond their empirical appearances. Obviously, many things that exist can be observed, but the existence of something does not depend on its observability. The most fundamental structures and mechanisms can often be established only analytically (Bhaskar, 1998, p. 41; Mingers, 2004, p. 93). According to CR, the intransitive reality – reality which is distinguished from the scientific discourses around it – is stratified into the *real*, *actual* and *empirical* domains. These are nested so that the real contains the actual, which contains the empirical. The domains allow

different epistemic access, which has profound methodological implications. The empirical domain can be accessed by direct observation, while the actual and real domains are investigated through retroductive theorizing that we will introduce below. The purpose of research is usually to uncover structures and mechanisms that account for relevant events, some of which are captured in empirical observations.

***Stratified ontology***

The domain of the real consists of objects, and mechanisms that arise from them. A structure is constituted by a group of component objects, which are interrelated in a specific configuration. A structure is an object itself because it expresses *emergent* properties that cannot be reduced to the individual components of the structure (Elder-Vass, 2007). For instance, an organisation is a structure that can have the capacity of producing aeroplanes, while none of its individual units or members has such a capacity alone. Component objects, such as organisational units in the example, are often internally structured in their turn. The constitutive associations that make an object/structure are called internal relations, whereas objects often have many external relations that do not affect their constitution or properties (Easton, 2010; Faulker and Runde 2013; Wynn and Williams, 2012). A collection of objects that expresses only the *resultant* properties of its parts is not a structure but an unstructured aggregate (Elder-Vass, 2005). Structures sustain mechanisms that account for causality and are the primary interest of scientific explanation. A mechanism can be understood as a capacity, that is to say, a possibility or tendency of what is likely to happen under certain conditions (Wynn and Williams, 2012, p. 791). Mechanisms are causal powers and must be activated for certain events to happen. Moreover, since objects/structures are

continuants, they can sustain causal powers across time and space (Easton, 2010; see also Faulkner and Runde, 2010).

In order to illustrate these abstract concepts, let us make some preliminary distinctions in the arrangements underpinning audiencemaking operations at the research site. To begin with, the telecommunications network infrastructure routinely generates a massive amount of individual CDR data tokens. These can be understood as relatively simple objects. Together, the CDRs constitute a data mass that may express emergent properties. The data are hence a potential structure, which can give rise to mechanisms that are relevant in audiencemaking. We call this candidate structure a 'data pool'. Our intention is then to investigate if the data pool has emergent properties that give rise to mechanisms shaping audiencemaking events and, ultimately, the audience product.

Events stem from the activation of mechanisms. It is worth emphasizing that the concept of event in CR is broad. For instance, "a bad year, a merger, a decision, a meeting, a conversation, or a handshake" can constitute an event that requires an explanation (Langley, 1999, p. 693; see also Wynn and Williams, 2012, p. 786). An event may happen only once or may be representative of a series of events that stem from the same mechanism. The kind of event to be explained depends on the research question that a study addresses. The domain of the *actual* contains all the events that take place, both those that are observable and those that remain unobserved, whereas the empirical domain covers only the events that are observable. The latter provide the starting point for critical realist theorizing about underlying structures and mechanisms.

***Retroductive reasoning***

Retroductive reasoning starts from an observed event and moves to theorizing the "hypothetical mechanisms that, if they existed, would generate or cause that which is to be explained" (Mingers, 2004, p. 94–95). The cause of an event is considered to be what makes a difference to its realisation. However, it is important to note that causal explanations are usually focused only on certain mechanics behind the event (Runde, 1998). It is often more interesting to analyse the event for specific features rather than whether or not it happens, or to try listing every possible mechanism involved. For instance, a press release is an event that is shaped by such factors as linguistic structures, public relations practices, managerial authority and a particular distribution channel. Yet, in this research we are interested in press releases and other events for the ways in which they contribute to the construction of an audience product. The same event can be accounted for in many different arguments, each focusing on a different aspect of the event and consequently providing a different kind of explanation.

Retroductive reasoning starts from empirical observations of an event. It then proceeds by analytically reconstructing mechanisms that would explain the event. The resulting explanation does not have to exhaust all aspects of the event, but it must be expressed in a way that allows the testing of its validity through further empirical studies. Theoretical explanations can compete when they result from attempts to capture the same structure or mechanism from different angles (Sayer 2000, p. 11), and they may eventually explain aspects of the structure that other theories ignore. However, the possibility of multiple theoretical explanations does not mean their equivalence. CR rejects a strong relativist position; its epistemic relativism does not

imply judgmental relativism (Mingers, 2004). Competing explanations can and should be compared, for the most accurate account of relevant causal mechanisms should have the highest explanatory power (Runde, 1998).

What makes discovery and validation difficult is that an activated mechanism may produce events that do not become observable in the empirical domain. There are often countervailing mechanisms that counteract or impede the manifestation of a mechanism to the observer. The regular observability of an event generated by a causal mechanism should therefore be considered a special case and not a prerequisite for a causal explanation (Runde, 1998, p. 153). The assessment of rival explanations should not depend on event regularities. Instead, a causal explanation must undergo a validation process that evaluates it according to different philosophical principles.

## Empirical Analysis

Our research site is a telecommunications operator that tries to turn its network subscribers into an advertising audience, that is, a product that can be sold to advertisers. The company was incorporated in 2006 after raising millions of euros in venture capital to launch a new kind of advertising platform. Operating as a mobile virtual network operator[2] but making money from advertising, the organisation has "the soul of commercial media, but the body and muscles of a telecoms operator", as one of the informants phrased it. Consumers could sign up for the service by providing a simple demographic profile and opting-in to receive advertisements on their mobile

[2] A mobile virtual network operator (MVNO) is a telecommunications operator that does not own a physical network infrastructure but leases it from another operator.

phones, while the company offered free voice call minutes and text messages in exchange.

### *Research design and empirical evidence*

Case study makes it possible to examine phenomena in their complexity, without reducing the object of research into just a few variables (Yin, 2009). This is an important advantage and makes the methodology compatible with a critical realist metatheory. CR supports intensive research that aims to identify and elaborate causal mechanisms rather than to quantify their efficacy (Easton, 2010; Wynn and Williams, 2012). Critical realist case studies typically answer *how* and *why* types of questions. They are suitable for unpacking circumstances in which the number of potentially relevant factors cannot be *a priori* narrowed down. An intensive case study like ours does not require a rigid explanatory framework to be fixed in advance, as its purpose is often to identify new explanatory mechanisms hidden from existing theories (Sayer, 2000).

The data collection took place during three-months' fieldwork using a variety of methods. One of the authors attended during regular working hours at the company headquarters, where he could constantly observe the 28 employees and directors located at the site. The staff consisted of experienced professionals in the fields of telecommunications, digital marketing, public relations, software development, business law, finance and management, organized into six teams responsible for different organisational functions. An observation log was constantly open on the observer's computer, allowing him to transcribe episodes as they unfolded and to avoid relying on his recollection after office hours. We define an episode as an uninterrupted

sequence of interactions that revolve around a common topic. Many (but not all) of the observed episodes can be understood as events that contributed to the effort to maintain a viable audience product.

At the beginning of the observation period, we had a broad interest in technology and business model innovation at the intersection of telecommunications and media industries. We quickly became sensitive to the role of audience measurement and, consequently, we narrowed down our focus to audiencemaking practices. These often drew on various measurement operations, tools and data. The observations were coded after the fieldwork period using a coding scheme derived from provisional explanatory ideas that emerged during the fieldwork. The purpose of the coding was instrumental rather than analytical. It allowed easy access to the episodes and gave proportions to the evidence, but the content and relationships between the codes are not central to the analysis. The process resulted in 689 episodes over 62 days of observation.

We interviewed 26 out of 28 people working at the research site; some informants were interviewed twice. The semi-structured interviews lasted from half to one hour and were based on a topical guide adjusted for each informant. The sessions were similar in structure, but the questions were tailored to the different roles covered by the informants and were designed to capitalize on recent developments at the research site. In order to map major events in the short corporate history and to achieve an insight into how the organisation presented itself to advertisers, we stored all the press releases and blog posts published on the company website. The observer also exploited serendipitous opportunities for gathering additional material. He stored documents and web pages, photographed events at the office, took screenshots from

information systems, and asked employees to provide examples of their instant messaging logs. Finally, we steered the fieldwork process on the basis of preliminary analysis. Every Sunday, the observer wrote an analytical memo (Walsh, 1998) reflecting upon the past week's efforts, identifying any problems or insights that should be addressed the following week. The summary of empirical evidence is presented in Table 1.

| Type of Evidence | Quantity | Details |
|---|---|---|
| Observation log | 62 days | 13 February 2009 – 15 May 2009 |
| Interviews (during the fieldwork period) | 34 | 26 different informants |
| Press releases | 26 | November 2006 – May 2010 |
| Blog posts (on the company website) | 60 | November 2006 – May 2010 |
| Intranet usage statistics | 335 days | July 2008 – May 2009 |
| Documents | 340 | Reports, intranet pages, etc. |
| Instant messaging logs | 59 | Conversations between employees |
| Photographs | 147 | Meetings, office events, etc. |
| **In-situ analysis** | | |
| Weekly summaries | 14 | One per observation week |
| Tailored interview guides | 34 | One per interview |

***Table 1 – The Types and Amount of Empirical Evidence***

In contrast to relatively clear methodological principles on how theories can be used as explanatory devices, refined and rejected, procedures for theory building are generally less formalized (Weick, 1995). Critical realism is particularly supportive in this respect, for it offers clear principles on how to theorize substantive phenomena (Bygstad, 2010; Easton, 2010; Wynn and Williams, 2012). The process starts with the identification and explanation of events which would contribute to answering the research question, and then moves to describing mechanisms and structures that are

expected to underpin those events. The former represent that which is to be explained (*explanandum*), while the latter provide the footing on which the explanation is built. A central part of critical realist analysis is retroductive reasoning, which moves from observations of events to hypotheses about mechanisms that could account for them. Finally, the hypothesized mechanisms need to be validated. Many critical realist scholars insist that the validation process should start within the study, but ultimately theoretical explanations need to be corroborated by other researchers and their independent investigations.

We conceive the retroductive identification of mechanisms as a process in which the researcher imaginatively fills the gaps between observed events with a causal account. The account explains what mechanism would produce the observed events and what structure would activate such a mechanism. For this purpose, we write an analytical narrative as a form of retroductive reasoning (Becker, 2007; Brewer, 2000). The narrative provides a medium in which it is possible to bring distinct observations together into an account informed by the critical realist metatheory. We start from specific audiencemaking events and reconstruct their connections with measurement data, gradually carving out three mechanisms operating at the research site. The weekly analytical memos made it possible for the process to be started already during the fieldwork. We allowed the past week's observations to inspire reflection and tentative explanations, which motivated attempts to fill gaps in provisional explanations during the following weeks. The resulting account is constructed to make relevant, empirically observed events intelligible by reconstructing their underlying causal mechanisms. The analytical rigor of the narrative is safeguarded by triangulation and two further guidelines. The variety of empirical evidence allowed us to triangulate observations and

therefore build confidence in our identification of important events and their features (Flick, 2004; Wynn and Williams, 2012). We also devised two guidelines to steer retroductive reasoning through our case. The guidelines helped to bring empirical evidence together systematically and to explore the meaning of most relevant tasks, operations and practices, while ignoring many fascinating but disparate episodes.

The first guideline is that the analysis should focus on events that are essential in terms of organisational survival. The viability of the enterprise would be decided by its success in attracting consumers and selling their attention to advertisers, that is, the execution of its novel business model. While the fieldwork deeply embedded us in the local setting and its shifting priorities, we identify relevant events as those that are necessary to sustain key business processes in the industrial context in which the enterprise operates. We call these *audiencemaking* events. Focusing on such events at the expense of others is consistent with the idea that retroductive reasoning does not have to account for all the structures and mechanisms present at the research site (Runde, 1998). The second guideline draws from the nature of the media industry and assumes that the importance of audience measurement has not vanished despite changes that are happening in the industry (Bermejo, 2009; Carr, 2008). The measurement of media consumption remains a central part of any effort to create a new kind of audience product. This further narrows our focus to the traces of measurement and analytical operations in audiencemaking events.

### *Audiencemaking events*

Let us start from a mundane episode that reveals a common feature in many work practices at the research site. The audience, either as a generic 'audience member'

or as aggregate 'members', is referred to, called upon and related with in daily operations. Such episodes occur frequently throughout the day and can be readily reported from the collected empirical evidence. The episodes designate events in which the new kind of audience is articulated along various dimensions. The audience does not come into being in a singular momentous event, but in a series of small episodes by which it is incrementally reinforced and shaped. For instance, in the following episode an employee (MCM) describes technological arrangements that are used to monitor the network subscribers (informants are represented by acronyms in the excerpts).

> *MCM discusses different member reporting models. At the moment there are three levels: ad hoc [manual], using dedicated reporting software and fully automatic. He talks also about the profiling of members for different countries. MCM says that a traditional operator does not care if the subscriber is away from the network for a few weeks, if the phone settings are correct, or if the phone model is up to date or not. While the operator may lose some revenue, it does not incur any costs. Therefore, it does not try to activate the subscriber. For us the consumers are the audience, for which we should have the connection.*
>
> *(Observation log, 16:15 on 24 March 2009)*

The excerpt shows how talk between employees routinely constructs network subscribers as members. We triangulated this observation between different kinds of episodes and documents, which confirmed that 'members' are discussed across the teams as well as in external communications. They represent the basic unit of the audience, and hence we call the instantiation of an audience member in organisational processes an audiencemaking event. The audience acquires its dimensions, is targeted with interventions and justified for various purposes by such events; in other words, the audience exists by virtue of continuous production of audiencemaking events.

People who subscribe to the mobile network are (obviously) never physically present, and it is from the information about their behaviour, rather than the human beings *per se*, that the audience is manufactured. The events include all kinds of interactions, operations and communications that occur in the company, from casual discussions and whiteboard scribblings to PowerPoint presentations, Excel spreadsheets and the release of marketing materials.

One might object that the audience is best understood as an interpretive construct in the context of organisational practices. However, this is simply not how members are experienced at the research site. The audience often react unpredictably to advertising and other corporate interventions. Some advertisements are even intended to build dialogue based on members' previous answers. Others get unsolicited responses. Feedback mechanisms are so common that audience reactions are regularly factored *a priori* into plans; the employees treat the member as an interactive entity, anticipating unexpected reactions. This can be observed in the ways in which employees harness a variety of reporting tools to get their work done. We identified 11 different systems for analysing and reporting from various sources of data. These include systems to track the delivery of advertising messages and member activity, to log and follow up the resolution of network issues and generic work orders, to create software development items and test cases, to measure the usage of company websites, or to monitor the company's reputation on the web. But, as we now proceed to argue, these tools would be of little support without the constant flow of fresh data.

***Data token object***

A digital telecommunications network makes a record of every click, call and message relayed through it, generating millions of records every day. These are known as call details records (CDRs). A network infrastructure needs to log traffic for various purposes, such as allowing the optimal allocation of resources, detecting and recovering from malfunctions, and identifying potentially harmful activity. The existence of such records is thus a structural pre-condition related to the functioning of the network infrastructure, rather than a decision by the company that harnesses the data to enable business model innovation. Therefore, while the records make the new kind of media business practicable, the genesis of CDR production falls outside the scope of the current investigation. The example below (taken from an unrelated specification document) illustrates the type of behavioural data that is generated by the telecommunications infrastructure[3].

> *097369D2D7372762D31080000000000000001;1;33668741168;3322208;6;20081101004923;20081101004923;20081101004923*
>
> *(CDR data token generated by a digital network infrastructure[4])*

The record captures the time, type, the sending and receiving ends of a network interaction, and a few technical details about the operation. The data token carries no reference to the social settings, intentions and activities that triggered the events that are captured in the data. Indeed, a CDR data token is a sort of receipt. It represents the delivery of an advertisement, or a network subscriber's response to it, as a text

[3] We are not allowed to reproduce an actual CDR from the research site.
[4] Advenage SMS Gateway Router 1.0 documentation

message. CDRs set the digital network infrastructure apart from traditional audience measurement arrangements in two ways. First, broadcasting advertising audiences used to be constructed from measurements of the reception of programme content, which can only indirectly reveal potential exposure to advertising that takes place during commercial breaks. Second, CDRs do not just measure exposure, but they also verify the individual responses to a specific advertisement.

The data is also extremely granular with respect to any practical purpose; CDRs merely turn ephemeral behavioural events into strings of alphanumeric characters that carry little meaningful content as such. The production of audience measurement data happens at this microscopic level of digital transmission receipts. The data record behaviour at a considerably higher resolution than previous audience measurement arrangements, well below the level of individual audience members. The raw data leaves open a massive gap between the tokens and a coherent audience product. Individual CDRs have none of the rich meanings the audience and its members carry in the context of organisational practices. A single reply to an advertising message, as captured by a data token, tells nothing organisationally relevant until it is combined with many others and is embedded into the context of a particular advertisement, campaign and a target group.

***Data-driven mechanisms in audiencemaking***

Next, we analyse several audiencemaking events and identify three mechanisms that enable an advertising audience to emerge from the data. The analysis builds toward a causal explanation of how advertising audiences are manufactured in the digital ecosystem. The identification and elaboration of mechanisms is also of key importance

in demonstrating whether the data pool is merely an aggregate of individual data tokens or constitutes a new kind of structure that expands the space of possibilities in the industry.

*Semantic closure mechanism*

During the fieldwork, we almost never saw raw data participating in organisational practices. The tokens are simply not practicable as such. As a whole, the data are voluminous and extremely detailed, suggesting that they could support a range of interpretations and insights. Yet, there is little actual information to work with in each individual data token, and turning their potential into facts about an audience is a far from trivial undertaking. Audiencemaking events that help to establish a new kind of audience product in the media market look quite different from the data tokens. For instance, an important event took place in August 2009, when a major industrial research firm confirmed some claims made by the company.

> Brands [advertisers] have been impressed with average campaign response rates of 25 percent. The richness of the interaction between Company's members and advertisers has also frequently been impressive. One example was a campaign organized by [Customer], which is a leading contact point for advice and guidance on bullying. The campaign was created to engage with 16- to 19-year-olds on this sensitive issue. Thirty-six percent of targeted members responded to the initial SMS [text message], and several of the responses revealed sensitive personal experiences and emotions. This type of engagement has convinced advertisers that mobile is a viable engagement medium for their target audiences.
>
> (Industrial analyst report, August 2009)

The event is notable in that an external agency supports the claims about the new kind of audience by circulating them through its report. The document specifically reiterates metrics that define the audience members by their behaviour. While the company had already put forward such claims on numerous other occasions, the analyst report effectively frames them as factual statements by a seemingly independent actor. Other similar behavioural constructions of the audience are found throughout the empirical evidence. For instance, the manager for advertising operations (BMA) described the product in his interview as follows:

> *BMA: Our [advertising] format is really good. It needs to be fine-tuned, but in general it is good: the response rate and all the behaviour we can generate – web traffic increases, coupon redeems and ROI [return on investment] for which it indeed culminates.*
>
> *(Interview of Business Manager, Advertising (BMA) on 13 May 2009)*

What makes it possible to conceive the audience as an interactive entity in the way that BMA does? The interactive characteristic contrasts with more traditional media. The construction of TV and radio audiences has historically revolved around the reception of media content by prescribed demographic segments, whereas the manager describes the new audience product as triggering and measuring behaviour. The shift from demographic to behavioural definition makes sense against the backdrop of the vastly improved measurability of behaviour. The essence of the new audience is not who it is but what it does. For instance, the rate at which the audience responds to advertising messages provides a good example of behavioural measures. It is referred to as the 'response rate' in the excerpt above, and, looking across our empirical evidence, the rate is one of the most important metrics the company uses to describe its audience.

The construction of the response rate metric presupposes suitable data and the means by which the data are combined together. Represented as a single number or a graph, the rate becomes part of the cognitive context for decision making and practical action. A concrete number can be pointed at, discussed and connected with many other events and measures, unlike an amorphous mass of CDRs. However, actual response rate readings could not form a foundation for other activities unless the mechanism by which they are produced remains stable over time. The rates are calculated by an algorithm that is embedded into the company's systems, filtering and combining data tokens according to a rigid procedure. The data are not coupled to a specific idea such as the response rate or any other metric that is brought into existence by programmatic operations. We observed a host of other metrics, including the number of active audience members, delivery of advertising messages and hyperlink clicks. These organisational metrics help to stabilize the focus on the inherently ambiguous audience. They render the audience product by producing its proportions on the specific dimensions of interest.

The data tokens are highly granular. They also capture a whole range of irrelevant, ambiguous and unexpected behavioural detail. For example, it cannot be decided, on the basis of data alone, if a repeated answer by the same member to an advertisement should be counted as one or two answers; or, what to do with a response to an advertisement that does not solicit any interaction. Such issues are not insignificant details. They indicate an important difference between a metric and the applications used to observe its actual readings. The response rate needs to be exactly the same irrespective of the application used to check its reading, which means that the

metric cannot be solely an artefact of the software application and its user interface. The actual readings are expected to change constantly (though not too much) in order to be perceived as a reliable reflection of behavioural patterns outside the system, but this needs to happen in the context of steadfastly coded procedures.

By 'semantic closure' we mean a stable way to interpret the data for a specific purpose, which is embedded and stabilized in technology. It then becomes taken for granted by relevant stakeholders. The automatic and continuous calculation of response rates is an example of a mechanism that provides a semantic closure on the data. The metrics become (and must be) black boxes for organisational practices. They hide their internal complexity, provide continuously updated readings, and remain stable over time. The metrics express these features consistently in all of their implementations. By stabilizing a specific procedure for interpreting data, the response rate algorithm allows a massive reduction of potential readings, collapsing them into one that becomes actual. It turns all but meaningless data into specific information about the audience.

*Pattern-finding mechanism*

The employees observe the metrics using a variety of reporting software applications. However, the applications do more than just generate the semantic closures that maintain the metrics. They are tools that allow user intervention by setting the parameters on how data is filtered, combined and represented in the context of organisational practices. Using the applications, the employees can mine the data for various kinds of patterns beyond the few stable metrics. Let us start from an event in which a certain aspect of the audience became suddenly unavailable. The following

excerpt depicts a situation in which a reporting system was perceived to fail in turning available data into information about the audience.

> *X1 comes over [to our table] and asks how should the large-scale operation on the member base be targeted. MCM and BMMA point out that the operation should be started immediately, because next week it might be too late. [...] X1 asks, which members are to be terminated. [...] MCM ponders what is reasonable and what is not. He points to the coffee table discussion in which it had been decided that the Member experience reporting tool will not be [immediately] updated. Resulting from this, we now lack adequate information for the decision.*
>
> *(Observation log, 18 February 2009)*

An outdated reporting application would hardly feel a problem if the data it represents do not matter. More specifically, the missing information appears against MCM's valid expectation of being able to elicit certain information from the data, which is based on his previous experiences on working with the tool. All in all, we identified five applications for retrieving, analysing and representing data on audience members (see Table 2). The applications enable employees to routinely represent aspects of the audience and its members, single out issues, and plan and execute both regular and *ad hoc* interventions. Many of the tools are used on a daily basis.

| System | Data source | Purpose |
|---|---|---|
| **Advertising reporting** | Network infrastructure | Reporting on advertising delivery and member interactions with advertisements |
| **Customer service system** | Call centre | The management of customer service requests |
| **Member experience reporting** | Network infrastructure | The analysis of subscriber behaviour in the network |
| **Web survey tool** | Online forms | A tool for creating and reporting web surveys |
| **Website traffic analysis** | Network infrastructure | The analysis of company website traffic |

***Table 2 – Applications Used to Monitor Network Subscribers as an Audience***

In contrast to the essentially rigid metrics, the logic of reporting applications is to enable multiple ways to arrange and summarize the voluminous data. The reporting applications are, first and foremost, user interfaces for querying multidimensional data. They enable employees to filter, combine and juxtapose data tokens, and to represent the results in tabular and visual forms. These representations often encapsulate organisational metrics discussed in the previous section. For instance, it is possible to compare the response rates for different advertisements in different geographical regions, between genders, and over time. The reporting applications help to uncover many patterns that may or may not be relevant, yet it is the data that ultimately set the boundaries and the possible paths for such explorations. The more data and dimensions a particular source offers, the more information a reporting application working with it can potentially reveal. The tools allow the situated judgement and inventiveness of employees to discover new avenues for making sense of the audience.

The pattern-finding mechanism is characterized by the role played by human operators, who need to devise strategies that could reveal more information from the data. Pattern-finding activities vary from mostly routinized activities to highly explorative attempts. In fact, we observed events that seem to express a different form of pattern-finding mechanism in operation. These events are associated with manually crafted analyses based on custom database queries and using statistical packages to analyse the output. Apparent problems in the network infrastructure, inexplicable member behaviour, or the needs of business development could motivate such a novel cut into the data. Also, potential information in the data simply drew interest from some

employees, who had consequently developed a habit of making casual data-mining exercises. The employees perceived and acted on the assumption that there is more information in the data than that which is being actualized by the current metrics and reporting applications.

Such exploratory opportunities are also harnessed by business development activities. Instead of precarious guesses about member behaviour and reactions to planned operations, it is sometimes possible to test assumptions by using reporting applications or by crafting a custom analysis. For instance, on one occasion it was necessary to dig deeper into the nature of member engagement with the advertisements. MCM, who was responsible for the member analytics, suggested studying the matter from the data. In a matter of hours he put together a graph depicting the speed of responses of different demographic groups. The visualization revealed interesting patterns beyond the aggregate response rate. For instance, it was found that the members either answer within a few minutes of the arrival of a message or are unlikely to engage the advertisement at all. Proposing such an analysis would have made little sense without the readily available data. The data pool provides a kind of laboratory environment where emerging ideas can be tested.

Learning from custom analyses also feeds back to the further development of measurement arrangements. Free-form explorations into the data can serve as initial steps for the development of new metrics and reporting applications. To summarize, the pattern-finding mechanism is made possible and boosted by the highly granular and comprehensive data generated by the digital network infrastructure. It also points to an interesting feature of the space of possibilities that the data open up. It is taken for

granted that there is potential information in the pool of data, but the amount of that potential information is unknown. The boundaries of pattern-finding are therefore *a priori* undefined, for it is not known in advance what can be done with the data.

The employees can query, tabulate and visualize patterns in the data using the reporting applications. The applications allow the activation of a pattern-finding mechanism. On the one hand, pattern-finding also provides a semantic closure on data tokens, but, on the other hand, the activation of the pattern-finding mechanism involves trying out and choosing between different semantic closures, not just reading a prescribed metric. Both the actual patterns and the ways to compile them can change, and, unlike the semantic closure mechanism, stability is not an overarching concern. The mechanism modulates between furthering established paths of semantic closure and the establishment of new ways to make sense of the data. The metrics and the use of reporting applications are the foundation for numerous reporting practices at the office.

*Framing mechanism*

The most generic reporting practice at the company is a weekly office meeting in which senior managers give brief updates on different aspects of the business to the staff. The meetings are held in the office lobby area as standing sessions without a formal decision-making function. For instance, we observed an event in which a senior manager (X3) asks about the size of the member base and tells briefly about the status of advertising sales.

> *X3 asks about the number of members. MCM answers that we have 75000 primary SIM card holders. X3 says that the number of top-ups is above*

*the budgeted and advertising sales are proceeding fairly well, even though achieving the budgeted sales will require very hard work. He continues to point out that the revenues of biggest media companies have dropped thirty per cent meaning that the market is really in a recession.*

*(Observation log, 10:00 on 9 March 2009)*

On an occasion such as the office meeting, the construction of an advertising audience becomes a largely interpretive exercise. The discussion about the overall audience size offers a good example. It may seem a simple, unambiguous number. MCM chooses to answer in terms of subscribers who use the company SIM card as their primary mobile phone subscription. This implies that that there are also other ways to count the number of members. For instance, the count would be different if it were reported as the number of people who hold a company SIM card. In a similar manner, the fact that sales are lagging behind targets is framed by the senior manager as fairly good by contrasting it to the current market conditions. The selection, timing and presentation of facts can matter just as much as information from the data. The office meeting was usually re-interpreted over lunch. In the lunch discussions, employees' views ranged from suggesting slightly different twists to the reported matters to debating what was the message that senior managers truly conveyed.

*People discuss some work-related matters over lunch. UED ponders that the tone in the office meeting was moderately positive. Others agree. HT jokes about running away to Bahamas with investors' money; AA continues that we are merely producing reports. Let's leave somebody behind to keep churning out the reports.*

*(Observation log, 12:56 on 9 March 2009)*

The comment about reports by AA is particularly revealing due to its inherent sarcasm. He acknowledges the importance of reports and reporting activities yet describes them as framing – 'we are merely producing reports'. AA thus suggests that reporting itself has become the focus of their work, not the things that are being reported. By framing, therefore, we mean the way in which the metrics and patterns observed in the data are brought to bear upon daily operations. The above comment is sarcastic because the employees are well aware that the mere practice of reporting is not enough. Behind the oral accounts put forward by senior managers at the office meetings, there are numerous reporting practices carried out in daily, weekly and monthly cycles in the organisation. In the context of such practices, employees selectively associate metrics and patterns found in the data with other sources of information, trends and objectives. The following interview excerpt shows how this occasionally went too far, generating reports which were too complex and which then required re-framing to again be useful.

> *HBD: X2 had one chap [in the local sales office] who compiled the statistics. And Operations team aggregated some other numbers and from these it was put together. [...] I was perhaps sometimes a little bit sceptical. We had sort of papers that incorporated 20 KPIs [key performance indicators]. For all those I told X3 and CEO that this is too complex. [...] In fact, I kept simplifying those numbers into Excel for myself even after we had the more sophisticated reporting, so that I could do the follow up [on member acquisition] compared to the earlier period.*
>
> *(Interview with Head of Brand and Design (HBD) team on 16 September 2009)*

Manually compiled PowerPoint presentations and Excel spreadsheets have a specific advantage over the pre-compiled metrics and the reporting applications. People

are able to select readings from different sources, combining and juxtaposing them with different tactics. In doing so, it is possible to strategically guide the interpretation of information to address issues from a specific perspective. There was often a lot of discussion on what a specific metric means for the task in hand, or what readings should be shown on a particular occasion or for specific material. For instance, it was not always clear how to count the number of active audience members against those lying dormant in the database. While this allows discretion and a degree of strategic ambiguity, without the data, metrics and reporting applications no credible reporting about the audience would have been possible.

In the three events described above, we perceive a mechanism that frames facts emerging from the data pool by virtue of the semantic closure and pattern-finding mechanisms. The purpose of the practical framing of facts is to more easily evoke certain interpretations while shunning others. At the same time, it produces new meaning that can be grasped only when the relationships between heterogeneous pieces of information are considered. Without such framing, the risk is that produced facts do not stand out or, even worse, are placed against an unfavourable background from the perspective of the company or an individual employee. The data pool alone is not enough to account for such a generic framing mechanism, which is activated, rather, at the encounter of interpretive agency and forms of aggregate data. The framing mechanism would merely produce an empty frame without the metrics, tabulations and data visualisations generated by the semantic closure and the pattern-finding mechanisms.

## Discussion

The new audience product is defined and maintained by the operation of semantic closure, pattern-finding and framing mechanisms that operate on the raw CDR data. The three mechanisms are nested so that an output from one feeds the other (see Appendix 1). This allows information about the audience to cascade through metrics, reporting applications and practices, becoming richer and more relevant for audiencemaking practices at every step. Table 3 summarizes the type of activating condition, observable entities and the typical operation of each mechanism.

| Mechanism | Activating condition | Observable entities | Typical operation |
|---|---|---|---|
| **Semantic closure** | The execution of program code | Metrics | Through stabilization of a metric, a continuous change can be observed from a fixed viewpoint |
| **Pattern-finding** | The use of reporting applications; custom database queries combined with the use of statistical packages | Tabulated and visual representations of aggregate data | Trying out and choosing between different ways to look at the data enables eliciting informative patterns |
| **Framing** | Reporting practices | Presentations, spreadsheets, verbal accounts etc. that contain representations of aggregate data | The production of more information by connecting the data to other data sources with respect to a broader context |

***Table 3 – Mechanisms***

Media companies have traditionally sold advertising space on the basis of the predicted amount of attention that a particular placement will attract, while the effective audience (those who actually saw the advertisement) used to be inferred *post hoc* from a sample of consumers participating in industrial audience measurement panels (Napoli, 2003). Our case study confirms and deepens the insight that the "institutionally effective audience" (Ettema and Whitney, 1994) is not made of people

but data. What cannot be measured cannot be verified to the advertisers and thereby cannot be part of the audience product. Against this background, the data generated by the digital network infrastructure introduces a major shift (Bermejo, 2009; Carr, 2008). The nexus of value creation shifts from obtaining valid and reliable samples of people's media consumption to analysing the audience from the extant data. Observing mobile phone users on the street would not help the company understand the audience because, paradoxical though this statement may seem, the audience is not out there but constructed from the data. In the following section, we elaborate the findings of retroductive analysis by theorizing a more generic mechanism and by identifying properties of the data pool. Finally, we will discuss the validity of the findings.

***Information actualization***

The advertising-funded telecommunications operator is, in certain respects, a relatively straightforward venture. The data pool offers a space of possibilities for the company to create a new kind of advertising platform with which to compete against both traditional advertising businesses and subscription-based network operators. A key assumption underpinning the venture is that the CDRs contain an informative potential, that can be extracted through automatic and manual elaborations, and then used to fuel audiencemaking operations. However, it is important to understand that valuable information is only potential in the data. It is something that can become expressed through certain events, or not. The data pool contains differences that are not *prima facie* meaningful (Bateson, 1972). We have shown in the analysis how, under certain conditions, these differences have an effect in the audiencemaking events (Bateson, 1972, p. 459; Kallinikos, 2006, p. 60–61; McKinney and Yoos, 2010). The relationship between the data as raw material and the audience as a product can be

understood through the Aristotelian dichotomy of potentiality versus actuality (Cohen, 2012).

Let us rely on a generally accepted understanding of actuality as the fulfilment of a potentiality, while potentiality indicates the possibility for something to happen, or come into being. The actual and potential are defined in relation to each other, one complementing the other. Aristotle argues in the *Metaphysics* that actuality stands to potentiality "as that which has been shaped out of some matter is to the matter from which it has been shaped" (1048b1-3 as in Cohen, 2012). Here, if we understand the data as the digital matter from which information is extracted, the three mechanisms constitute a set of *information actualization mechanisms*. Information actualization describes various ways to exploit the new space of possibilities that exists by virtue of pooling vast amounts of digital data.

The idea of information as actualized potential is analogous to the classic marble statue example. Russell (1994, p. 180) writes "'a block of marble is a potential statue' means 'from a block of marble, by suitable acts, a statue is produced.'" The block of marble (data) neither determines the existence of the statue nor its shape (information), but it is equally true that the statue could not appear out of nothing. The potential does not exist in material alone, but requires the availability of means to transform the material into something else. It takes a combination of suitable skills, actions and material for something to happen or come into being.

***Properties of the data pool structure***

The foundations of the semantic closure and pattern-finding mechanisms we have identified lie in the structural properties of the data pool. The practical conditions for their emergence stand in the sheer amount of data and the technological capacity to simultaneously filter and combine a large number of tokens. We identify three properties that define the data pool structure: the *comprehensive*, *granular* and *unbounded* characteristics of the data pool.

To begin with, the digital data tokens matter because the digital network infrastructure automates much of the data collection. In traditional media, this is done by separate measurement devices distributed to a small subset of consumers. The collected data is then limited to carefully planned samples geared to predefined purposes, whereas in the present digital ecosystem the behaviour of the whole user base is captured implicitly by the infrastructure. There is no need to distribute and maintain the expensive metering devices. Importantly, the massive amount of data generated by the digital infrastructure is not a sample but the census of the activity in the network. The data pool can be said to be a *comprehensive* collection of user behaviours.

The digital network infrastructure not only automates the data collection, but also generates records which are qualitatively different, as compared to earlier audience measurement arrangements. CDRs were not designed for audiencemaking purposes. They dissolve media use into discrete clicks and messages. It is from the pool of such extremely *granular* behavioural traces that meaningful behavioural patterns have to be reassembled by recourse to analytic operations (Kallinikos, Aaltonen and

Marton, 2013). If the data collection was earlier framed as surveying predefined consumer segments and categories, those have to be now produced *a posteriori* from the extant data. The meaning lost in the extreme granularity of the data is, however, compensated by the vastly expanded opportunities to aggregate, align and juxtapose digital data tokens against each other (Kallinikos, 2006; Kallinikos, Aaltonen and Marton, 2013).

Finally, the individual data tokens represent ephemeral behavioural episodes, which give them a "use-agnostic" character (Kallinikos, 2012). The data are loosely coupled with the uses to which they are actually put and may not immediately seem able to answer any relevant question. They exist as an open-ended potential, to be explored in a variety of ways and to different ends. Importantly, the pool of agnostic data tokens leaves the boundaries of such explorations open and undefined. This makes the space of possibilities emerging from such data look characteristically *unbounded*. What can be done with the data depends on the availability and activation of specific information actualizations mechanisms.

Table 4 summarizes the three properties of digital data in the case. The properties are hardly idiosyncratic to the case, but we acknowledge that other cases may also exhibit other properties (Ekbia, 2009; Faulkner and Runde, 2010; Kallinikos, Aaltonen and Marton, 2013; Yoo, Henfridsson and Lyytinen, 2010). While comprehensiveness and unboundedness are attributable only to the data pool as a whole, granularity could be understood as a property of the individual data token object. The former two are thus emergent properties (Elder-Vass, 2005, 2007); they appear as large amounts of data tokens and are managed in relation to each other. The

presence of emergent properties suggests that the data pool is a new kind of structure and should not be considered just a heap of data. It has causal powers that support the activation of the mechanisms we have found through the analysis of empirical evidence.

| Property | Type | Description |
|---|---|---|
| Comprehensive | Emergent | The data is the census of activity in the system (not a sample) |
| Granular | Resultant | The data tokens break a referent reality into meaningless behavioural episodes |
| Unbounded | Emergent | The boundaries of data-driven understanding are not known in advance |

***Table 4 – The Properties of Data Pool Structure***

Let us briefly qualify the three properties and explain why we think they are either emergent or resultant properties (Elder-Vass, 2007). To begin with, comprehensiveness cannot obviously be attributed to an individual data token. It results from the collection of the totality of behavioural events in the network and, unlike a sample, allows individual interaction with each member. The case is different with regard to granularity, which, in our case, concerns the resolution at which people's media use is recorded. A data token represents a single member interaction and, in this respect, granularity is a resultant attribute of individual objects in the data pool. Nevertheless, a highly granular pool of data tokens enables the data to be explored by many more combinations than a less granular pool of data would allow. The third property, unboundedness, and the other two properties above, are interrelated. The potential of the data to inform about many unforeseen issues would be limited without the comprehensiveness and granularity of the data. It is the combination of breadth (comprehensiveness) and resolution (granularity) that explode the number of potential

questions that can be asked from the data. Unboundedness is thus an emergent property.

### *The validity of the findings*

The three mechanisms described in this study are candidates for causal explanations of the observed events. The critical realist metatheory requires the results to be presented so that they can be tested against alternative hypotheses, and it has been argued that studies should include an assessment of the identified mechanisms against other possible explanations (Bygstad, 2010; Runde, 1998; Wynn and Williams, 2012). We first consider an alternative kind of explanation to the audiencemaking events and then discuss the analysis against a set of evaluation criteria for causal explanations.

A possible alternative explanation could be based on the assumption that the properties of digital data have no significant impact on audiencemaking events and, consequently, on the audience sold by the company. One could try to argue that it is possible to understand the audience in terms of the coalescing of interpretive acts. The response rate and other characteristics of the audience product could be analysed as choices made by the actors and not as outcomes shaped by the mechanisms that emerge from the digital data. The alternative explanation would then centre on negotiations and interpretations in the process of constructing the audience. However, important aspects of the case escape this kind of explanation. The audience members are found to behave in unexpected ways in the data; they surprise employees and shape their plans and expectations. Furthermore, the occasional inability to turn data into information would not hinder action if the data pool were not making a difference to organisational

practices. The alternative explanation limited to the interpretive dimension of organisational practices would fail to recognize the specific ways in which the data enabled and constrained the construction of the audience.

Runde (1998) proposes four principles for evaluating a retroductive causal explanation. A causal hypothesis is considered plausible and well-formed if the candidate mechanism: is taking part in the situation where the observed consequence occurred; is a plausible cause of an event that needs an explanation; is deemed sufficient to cause the aspect of the event under scrutiny; expresses a degree of causal depth (it has explanatory power). In regard to the first principle, the three structural properties of the data pool and the three mechanisms are clearly implicated in audiencemaking events. Second, the reactions and interpretations with respect to the data are events that warrant an explanation, since they are critical to the success of the company. We have shown how important aspects of the events could not be understood without unpacking the role that the data pool plays in their unfolding. Third, our explanation is sufficient in that we retroduced a set of related mechanisms that, if they were real, would explain why the observed events construct the audience in the way they did. We aimed to postulate only the structures, mechanisms and powers that it is necessary to take into account at the level of abstraction at which we are developing our argument. The explanation does not exclude other intervening or countervailing causal powers. For instance, we have identified the presence of an interpretive element contributing to the framing mechanisms that is involved in constructing the new kind of audience. Fourth, the argument has causal depth. It explains how an advertising audience is constructed in the digital ecosystem by reference to specific mechanisms and the data pool structure.

## Conclusions

In this article, we have demonstrated the use of critical realism for studying the production of data-driven products and services. The argument was substantiated by analysing how a telecommunications operator transforms agnostic data from a network infrastructure into valuable information about a new kind of advertising audience. Critical realism helped to pin down audiencemaking events against a relevant industrial background and then analyse how the audience is manufactured from the data. The findings are based on a single case study, but our contribution toward understanding the mechanisms of information actualization could be broadly validated.

The findings are relevant and timely. Information systems do not just store, process and transfer data, but they also generate vast amounts of new data. New data may have initially been generated for only peripheral uses (such as maintaining the network itself), but they are also increasingly recognized as raw material for new products and services. Indeed, products such as advertising audiences, securities, insurances and many kinds of ratings could be called 'data-based' rather than data-driven, for they are made out of data (Redman, 2008). Recently, there has been a lot of excitement and discussion about the opportunities of 'big' and 'open' data. In several ways, the research site represents many of those organisations that execute novel business models around what is perhaps vaguely termed Big Data (Boyd and Crawford, 2012).

Whether data-based business opportunities can be realised depends on an organisational capability to harness the potential embedded in newly available digital

data. Many organisations are at a loss with these opportunities. They either sit unknowingly on top of an enormous resource or lose themselves in the morass of meaningless analytics (Aaltonen, 2012; Day, 2003). Building metrics and developing reporting tools and practices are seldom perceived as the most interesting activities in an office, but understanding them is critically important to an increasing number of businesses. The data has no value without the arrangements that can realize its potential; our study is a concrete example how those arrangements can be studied and offers a set of mechanisms as a starting point.

More generally, our study differs from the body of IS literature in which computing is "conceptualized as a discrete symbolic representations of something in the *real* world" (Yoo 2010, p. 218). The individual data tokens may be understood to represent actions of flesh-and-blood human beings, but the audience does not have such a clear, external referent. The aggregate of digital data (what we define as the data pool) is real matter with emergent properties. The product is literally manufactured from such raw digital material. Supported by a critical realist metatheory, IS scholars can be at the forefront of explaining the transition from the mere processing (or *reading* as in Zuboff, 1988; Kallinikos, 1999) of technological representations to new socio-technical configurations that involve the construction of new products and forms of value creation on digital data. Wikipedia and open source software development are good examples (Aaltonen and Kallinikos, 2013; Benkler, 2006), but there are many others.

We believe that digital materiality needs to be studied intensively, that is, by theorizing emergent properties specific to the digital ecosystem. While we are sympathetic to the agenda set forth by Leonardi (2010), the analysis of digital

materiality as emergent properties and mechanisms raises issues with respect to the definition of materiality as "practical instantiation of theoretical ideas" and "what is significant in the explanation of a given context" (Leonardi, 2010). These two definitions provide useful perspectives, but they exclude certain aspects regarding how the digital ecosystem matters in business. Digital data, in the form of structures such as a data pool, do more than just instantiate theoretical ideas. Ideas often require material underpinnings to be conceivable in practical terms. There is no reason why ideas should pre-exist materiality – some may, but the opposite situation can also exist. Working hands-on with materials stimulates curiosity and imagination, making it possible to develop new ideas (Dourish, 2001). We have shown throughout our study that a data pool defines a space of possibilities. It is the matter within which a number of work efforts are imagined, conceived and executed. Our theorizing generally agrees with Leonardi's second definition, but it is important to point out that the emergent properties of digital data are not straightforwardly read off from empirical observations. Understanding 'material' as that which matters for a given activity is a good starting point (cf. Latour, 1999). However, we also need robust conceptual tools to analyse how generic attributes of the digital ecosystem matter in specific industries and organisational settings.

# Governing *PatientsLikeMe*: information production in an open, distributed and data-based research network

**Niccolò Tempini, LSE**

**Abstract**

In this paper, I set out to understand the specific conditions shaping the production of information through social media networks, with a focus on the role of data structures. Many organisations develop social media networks with the aim of engaging wide social groups in the production of information that fuels their processes. This effort appears to crucially depend on complex data structures that afford the organisation to connect and collect data from myriad local contexts and actors. One such organisation, *PatientsLikeMe* develops 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. Once contextualized in this case, the question on 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. Through an observational case study, I show 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. To explain how the organisation reacted to these challenges, I introduce the concept of *information cultivation*, understood as an organisational strategy characteristic of specific data

collection architectures exploiting open, distributed and data-based arrangements (such as social media). With the concept of *social denomination* I try to capture the form of governance of the patient audience associated to information cultivation efforts. The study adds new insights to previous research efforts regarding how information stems from data that translate across contexts in variably standardized forms, and discusses some of the social consequences of social media models for knowledge making.

## Introduction

New organisational forms have emerged in association with the widespread diffusion of web and social media technologies across the social fabric (Howe, 2008; Shirky, 2008, 2010). Organisations 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). Typically these systems routinely produce information from the data that users generate while dealing with the matters of their own lives. As suggested by Howe (2008), the capillary reach of these networks might better capture the ephemeral but valuable knowledge of diverse and distributed local contexts, which tends to escape universal models and covering law explanations (Hayek, 1945). New socio-technical configurations powered by social media seem to capture and repurpose the trivia from users' everyday living into data (boyd and Crawford, 2012; boyd and Ellison, 2008; Kallinikos and Tempini, 2011; Mayer-Schönberger and Cukier, 2013), making them amenable to inclusion in networks of economic relations. Nonetheless, using 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 organisation *PatientsLikeMe*, based in Cambridge, Massachusetts, a well-known venture exploiting the possibilities offered by social media technology to set itself at the crossroads of the pharmaceutical and health services industries, patient organisations and advocacy networks, care communities, and Internet research. The for-profit company, founded in 2004, has developed an ad-free social networking site whereby patients can connect with each other as they collect self-reported medical data.[5] The research team exploits the collected data for scientific and commercial medical research purposes, attempting to evolve the model of business intelligence through social data collection and analysis to meet the requirements of medical research standards. Authors have welcomed these innovative forms of scientific enterprise (Shirky, 2010; Topol, 2012), anticipating great innovation and disruption might be unleashed if research leaves the artificial setting of the laboratory and the clinical hospital and reaches out into the real world. To date, the researchers working on this network have produced 37 scientific publications, based on data contributed by

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

more than 220,000 patients. Research outputs include peer-reviewed articles, conference papers, reports, editorials, and others. Contributions have covered a broad range of subjects, with a few remarkable results. 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 (PD) patient population by comparing it to another patient population dealing with a chronic progressive neurological disorder (ALS), 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).

The organisation styles itself at the same time as an all-encompassing platform for the organisation of patient sociality and advocacy, aiming to become the social media network of choice where relationships between patients, clinical professionals, healthcare providers, pharmaceutical companies, patient organisations and NGOs are discussed or intermediated. In this sense, *PatientsLikeMe* differs from patient and evidence-based activism organisations (Epstein, 2008; Rabeharisoa *et al.,* 2013), pioneering a new kind of intermediary. Critically 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 organisation's business model is designed, once embedded in the incumbent network of economic relations most of the research the network has

produced depended on funding from related commercial research projects. The position of this organisation as a novel actor in a saturated and resistant to change institutional landscape, trying to re-open the boundaries of the scientific enterprise (Gieryn, 1983), is certainly worthy of evaluation. But, these considerations are beyond the scope of this article. Instead, the point that I put forth for discussion is political and social yet relies on a deeper level of investigation than is usual in arguments on sector structures, discourses and agendas. At a time when massive communication networks are entering various spheres of public life, coming to intermediate the social at all levels, it is necessary to develop a thorough understanding of the roles information infrastructures and their data configurations play in organising social projects (Agre, 1992; Bowker and Star, 1999; Star and Lampland, 2009) such as, in this case, medical research.

At the center of *PatientsLikeMe*'s innovative approach to medical research, the "raw matter", as it were, that is worked upon in the making of research, stand the data that the network routinely collects from patients. *PatientsLikeMe* engenders *'data-based production'* through social media technology (Aaltonen and Tempini, 2014), depending fully on the collection of social (health) data generated by the patients through the technological infrastructure the organisation controls, and on the processing of the data for the production of information essential for the services the organisation offers. For understanding an innovative organisational form such as that represented by *PatientsLikeMe*, it is critical to explain the conditions that shape the production of information out of data. First, data collection through social media means that the researchers do not learn about the patients, their experiences and their health situations in any other way than through the social data – what patients write or do in this environment. The social media infrastructure that the organisation develops on a

continual basis, in the fast-paced fashion of web-based development, 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) with the fundamental contribution of the patients – as data entry operators and immediate observers of medical realities.

At the core of the infrastructure, data structures are *'gateway technologies'* (Ribes and Bowker, 2009:201), translating knowledge between the organisation and a myriad of local contexts. With such premises – of organising research through social media and the massive involvement of an open and distributed user base – there are pressing questions to answer. *First, 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? Second, how is the patient user base governed so as to select and encourage behavior that supports the fulfillment of said requirements?* The data are indeed generated by patients from the most disparate contexts. This is the paramount challenge for contemporary organisations that produce information through unconventional, open and distributed data collection arrangements.

To answer these questions, we must explain the often-invisible work processes and devices that make data comparable and translatable across contexts (Star and Lampland, 2009; Star, 1983, 1986). I claim that the literature studying social media phenomena from innovation and organisational perspectives has perhaps not been sufficiently concerned with the making of data structures. Research has focused on how social media, with their power to erode spatial and temporal distance, afford new

organisation forms for knowledge production (Treem and Leonardi, 2012), facilitating exchanges within or beyond organisational boundaries (Majchrzak *et al.*, 2013), or supporting the generative liveliness of seemingly self-organized online communities (Faraj *et al.*, 2011). Analyses have emphasized how these networks link users, content, and combinations of the two (Treem and Leonardi, 2012), but have not unpacked the issue of the role data structures and models, or other specific technological structures, play in the construction of these connections. More research is needed if we want to understand how organisations controlling social media networks set about governing their user base, and the conditions characterizing this endeavor.

Classification systems embed labels and numbers representing and ordering people, their interrelationships and their life contexts (Bowker and Star, 1999; Timmermans and Berg, 2003; Timmermans *et al.*, 1998). 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) by which similarities and differences between phenomena are made explicit and real. To represent knowledge in data structures means to articulate in practice what Leonelli, in the case of bio-ontologies, calls *'classificatory theories'* (Leonelli, 2012). 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'* (Leonelli, 2012:58). Despite differences with the case of bio-ontologies, I hold to this perspective when investigating the data structures embedded in the *PatientsLikeMe* social media network.[6]

However, issues of ontological representation are not simply a theoretical dispute. Instead, they are ground 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. Research on the development of distributed information infrastructures in science has explored the challenges that scientists and developers face in coordinating work across contexts and ensuring data interoperability (e.g. Millerand and Bowker, 2009; Ribes and Bowker, 2009; Ribes and Jackson, 2013). Diversity of contexts, organisational structures, installed bases (Hanseth and Monteiro, 1996) and classification structures all affected the continuity and comparability of data collection. Knowledge representation and embedment in information infrastructures is

[6] Most notably, as it becomes apparent through my empirical narrative, the data structures in *PatientsLikeMe* are subject to systematic scrutiny only between the organisation and the research partners, as their limited visibility from outside – embedded in the workings of the system – does not facilitate further warrant.

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).

The issues, which this literature has thoroughly explored, are repurposed, yet perhaps made more complex, in the context of research data collection through social media. When the participant patient community is inherently open, distributed, diverse and yet undefined, it becomes particularly critical to find a balanced configuration of the data structures powering the social media information infrastructure. In *PatientsLikeMe*, data collection is performed at all times, from virtually anywhere, by patients that are not directly known, briefed, or cross-checked in any particular way. Data that patients report are aggregated and displayed across the site for the benefit of both patients and researchers. While the patients collaborate on the platform for multiple and different reasons – including seeking for a cure and socialization, solidarity and friendship – the researchers try to encourage particular data collection behaviors. In this purview, it is clearly of paramount importance to understand whether specific configurations of data structures differently perform as efficient organisational devices, *'semantic gateway technologies'* (Ribes and Bowker 2009:201) enabling communication and coordination between different patients and life contexts.

Building on Bateson's definition of information as *'difference that makes a difference'* (Bateson, 1972),[7] Jacob (2004) compares between systems of categorization

[7] Acknowledging that I am not doing justice to the work of such a complex thinker here, it is sufficient to understand that information is an event (the difference in the making) that depends on a phenomenon (a difference that is marked). The information

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 organisation. 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. Although only sketched, this inverse relationship between semantic context and local context flexibility offers a preliminary framework for making sense of information production in the *PatientsLikeMe* context. In this perspective, we must contextualize our interest and ask, *how do data structures capture information from one context and transfer it to another?* And, *what factors shape the amount of information that can be expressed by data collected through an open, distributed network?*

In this paper, I answer to these and the earlier questions by explaining the challenges that *PatientsLikeMe* faces in developing a system aimed at maximizing the amount of information that can by produced from the data collected and aggregated

---

expressed by a piece of data (a marked difference) saying, for instance, that '*x* is A' is a function of what the system of signs used to mark the difference tells us that *x* is not. A system that differentiates more carries more information. An example is the twenty-six-letter English alphabet as opposed to the Chinese ideograph system: *'The actual letter excludes (i.e., eliminates by restraint) twenty-five alternatives. In comparison with an English letter, a Chinese ideograph would have excluded several thousand alternatives. We say, therefore that the Chinese ideograph carries more information than the letter'* (Bateson, 1972:408).

with the patients' contribution. I show through detailed empirical evidence that the challenge is often paradoxical. In an open and distributed data collection network trying to increase the amount of information produced by increasing the degree of structure in the data models (thereby increasing information through an increase in data specificity) often comes at the expense of decreasing user engagement (thereby decreasing information through a decrease in data scale). A more complex, restrictive or time-consuming system is typically used by fewer people. Conversely, aiming at high levels of patient engagement often comes at the expense of data specificity.

In *PatientsLikeMe*, this inverse relationship actively shaped the development of the system and its data models. Both dimensions (data scale and specificity) influence the informative potential – the potential to produce information – of the data that the organisation cultivates. Aim of this paper is to explain this paradoxical relationship by theorizing about two mechanisms of *information cultivation*. Highlighting these phenomena as characteristic of the operations of organisations that set out to produce knowledge through social media-based arrangements, I proceed to draw some of the major implications of the use of social media technology for the governance of publics and the production of real-world knowledge. Before I move on to the next section, however, I would like to ease the reader's journey into the empirical setting by offering an immediate depiction of the phenomena I have just abstractly delineated. The following fictional vignette (inspired by reported evidence) intuitively exemplifies an instance of the problems I am going to analyze. The vignette shows the radical, yet at the same time mundane, character of the dilemmas that have characterized the development of the *PatientsLikeMe* system for *information cultivation*.

*Margaret is 63 and suffers from arthritis. She was diagnosed with osteoarthritis a couple of years ago, but symptoms had appeared a few years earlier. Margaret learnt a few days ago about a website where she could track her health and meet other people. She wants to give it a try. She creates her account, fills in her personal details, and goes on to input data about her health situation. The system now asks her to specify the condition she is suffering from. She types in the search box 'arthritis', but the condition does not appear in the results. Instead, the system suggests she may have 'osteoarthritis', 'rheumatoid arthritis' or 'psoriatic arthritis', among others. Margaret doesn't know what to choose. She doesn't remember that she has osteoarthritis, the most common form of arthritis. As far as she remembers, she's always had 'arthritis'. Margaret is confused and decides to abandon the task for the time being. She stops filling in the questionnaire. She goes on to the forum area of the website and reads a few threads. Eventually, in the next few days, she will forget about the questionnaire and will log on to the platform only periodically, just for a bit of interaction in the forums. The system will not know her condition, and it will be very hard to engage Margaret in other data collection tasks.*

*Sandra is a member of the integrity team that supervises the index of conditions, treatments, symptoms, and other medical entities recorded in the system, to which users can link their data. She has been struggling with the problem of modeling arthritis. Back when they allowed patients to identify themselves as having arthritis in their profiles, many patients did so. Truth is, all of these patients have a subtype of arthritis, but many of them don't really know which, either because they don't have much medical knowledge, or because they have just forgotten over time. Sandra saw that she was not getting good data, with all those patients identifying generic arthritis as their condition in their profile. Of course, a lot of patients were recording the condition and related symptoms and*

*treatments, which is usually a good thing, as more patients and data mean more material for research. However, all those data were of very little value, as Sandra couldn't tell whether an experience was of rheumatoid arthritis or psoriatic arthritis. Therefore, she decided to take action. She disallowed arthritis as a generic term and required patients to either find their subtype, or not record that they had arthritis. Since then, the number of patients stating their condition has dropped significantly, but at least now the context surrounding items such as particular symptoms in the patient profile is a little bit clearer. Based on her own experience of curating the PatientsLikeMe system, Sandra prefers - in this case - to collect less data because otherwise the data would not be sufficiently specific and meaningful.*

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 organisation, 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 organisations and Internet medical research.

## Methodology and research design

For 26 weeks – from September 2011 to April 2012 – I conducted an observational case study (Yin, 2009) at the headquarters of *PatientsLikeMe*, based in Cambridge, Massachusetts. The organisation has been developing a health-based social

networking site for connecting patients of all diseases. 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 representing the organisation at conferences, meetings or conference calls. I acted as a regular member of the staff working on the development of the social media information infrastructure.

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 on developments in my observations. I had extensive access to organisational resources, and I was able to obtain more resources when needed. The flexible nature of my participation in the organisation 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 had developed tentative interpretations of the phenomena I had been observing through constant analysis and interpretation. In my time off-site (evenings, weekends), I reviewed and further integrated my notes, trying to reflect on my main findings from recent developments and to explain the events I had observed (Mingers, 2004; Sayer, 2000). I used retroductive reasoning as a technique for theorizing on the kinds of mechanisms relating activities, people, and technology in the unfolding of events. Retroduction is a process by which, 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 empirical manifestation (Runde, 1998). I deemed relevance 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 interpretations as well as narrations of events as I experienced them, 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 the 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 organisational 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 organisation 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 5, I present a census of the data I collected or generated during my fieldwork.

| **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 reflections)** | 665 |
| **Meetings (with minutes)** | 128 |
| **E-mail exchanges** | 1670 |

***Table 5 – Data generated on site***

## 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 organisations 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 organisation conducts and publishes. A main, overarching concern for the organisation 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 organisation 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 their 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 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.[11] 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

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

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 symptom,[12] 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.

[12] Symptom severities are captured along a NMMS (none, mild, moderate, severe) scale.

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.[13] 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'* (see Kallinikos and Tempini, 2014:830; Alaimo, 2014).

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

[13] See,Chapter 5 'On tuberculosis and trajectories' in Bowker and Star (1999), for an stimulating discussion on the relationship between health measurements, and biography.

phenomena. Exploiting the continuous updatability of Web-based applications, the organisation 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[14] 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.[15] As a researcher at the organisation 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.*

[14] As of September 2014.
[15] 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 express 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']*.'*[16] 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.[17]

***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

[16] 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).
[17] 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.

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 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.[18] 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

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

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.[19]

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 organisation's 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.[20]

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

[19] 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.
[20] E.g. free form, pill, vial and etcetera.

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 organisation realised 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 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 add no more 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).[21] Figures 1 and 2 descriptively represent this trade-off in a simplified fashion, by showing

[21] 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').

two alternative set-ups of condition categories and the consequent effects of the scale of data collection.

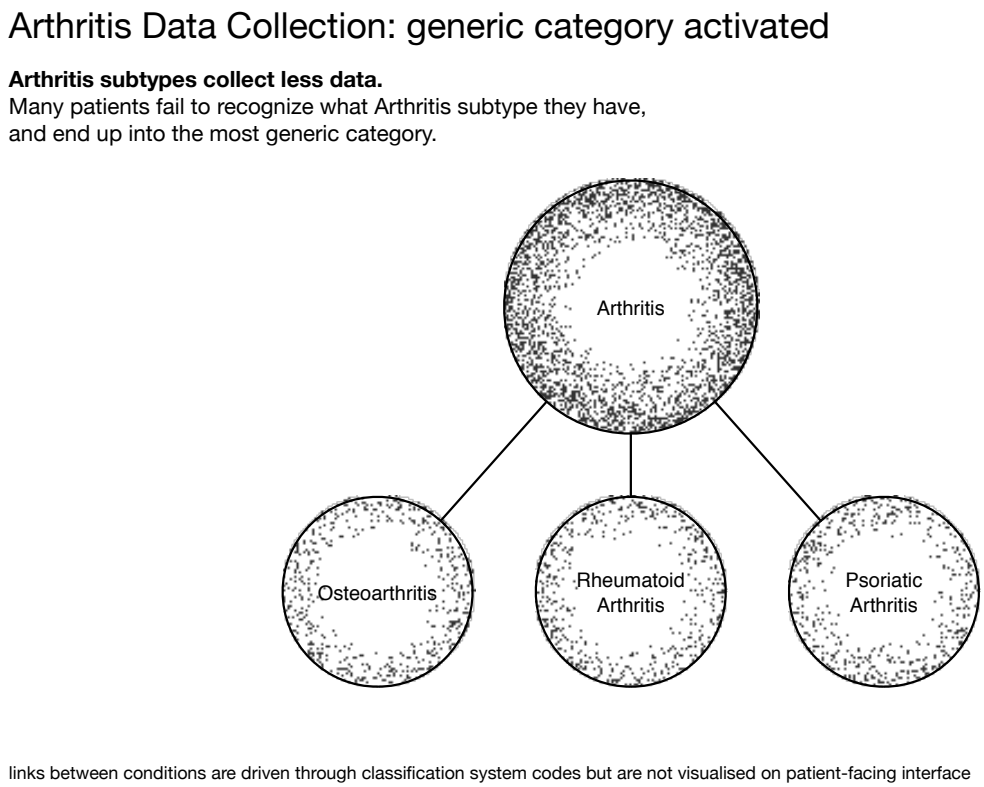



***Figure 1: Data collection including generic Arthritis***

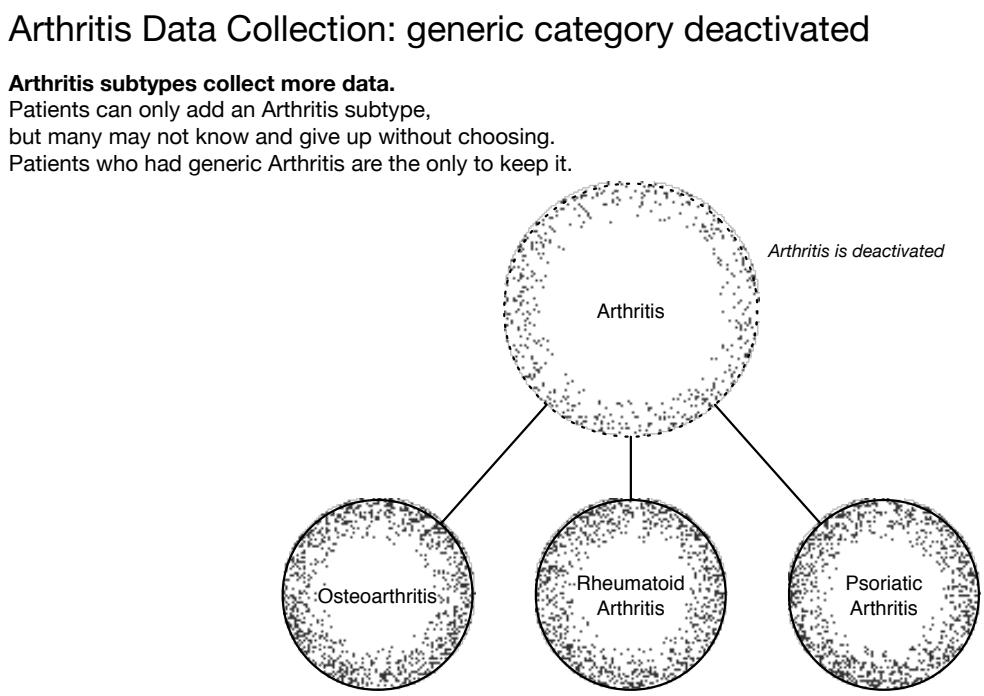



***Figure 2: Data collection excluding Arthritis***

As the organisation tailors the system in order to produce more information, both data scale and semantic context shape system development efforts. Obviously, the organisation 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 During the observation period, the staff often discussed how to gauge the information potential – the potential to produce‘ *During the observatio*. An executive described this as ‘ *An executive described this asure of the value of the fundamental underlying database.b* 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; itulation, 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*

***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 patientsa specile in two ways. First, upon account creation, the patients profile is attributed five generic symptoms deemed applicable to any patient experience.[33] Second, the system automatically adds a number of condition-specific symptoms to the

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

patients two ways. Fin the patients add a condition to their profile.[34] 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 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

[34] 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.

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.[35] 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 investigate whether the definition provided by the patients does in fact describe a symptom.[36] 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

[35] This is also possible for other medical entities such as conditions and treatments.
[36] 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.

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 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.[37]

Through this open, participatory data collection process that recognizes the patient a role of observer and operator (see also Kallinikos and Tempini, 2014), 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.[38] 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 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

[37] 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.

[38] 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). 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 realisation of a disease*' (Scheuermann *et al.*, 2009). For instance, the clinician does not directly observe a symptom such as a headache. Only the patient has access to the phenomenon.

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).[39] 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, and importantly, the researchers found that many symptom terms express *'Other kinds of anomalies, however, are lessnular terms than the UMLS the UMLS* (Arnott-Smith and Wicks,

[39] 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.

2008:685). 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 organisation 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,

> *Thereranslate and aggregate related yet different patient symptom definitions when it became necessary for research purposes. This labor-intensive mapping operation ngrequiring reseaays.* [he *If someone puts in gregate related yet different patient symptom definitions wh*[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 ]person who has had a symptom that h*[symptom definition] *that the patient actually told us about in their own wordsmptom that hit on the same Me*[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 organisations 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, from within the organisation. 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 organisation tries to engage the general public in information production. In the first instance, we observe that organisational 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 organisation 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 remark 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) are often related in paradox. 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 the appendix, 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.

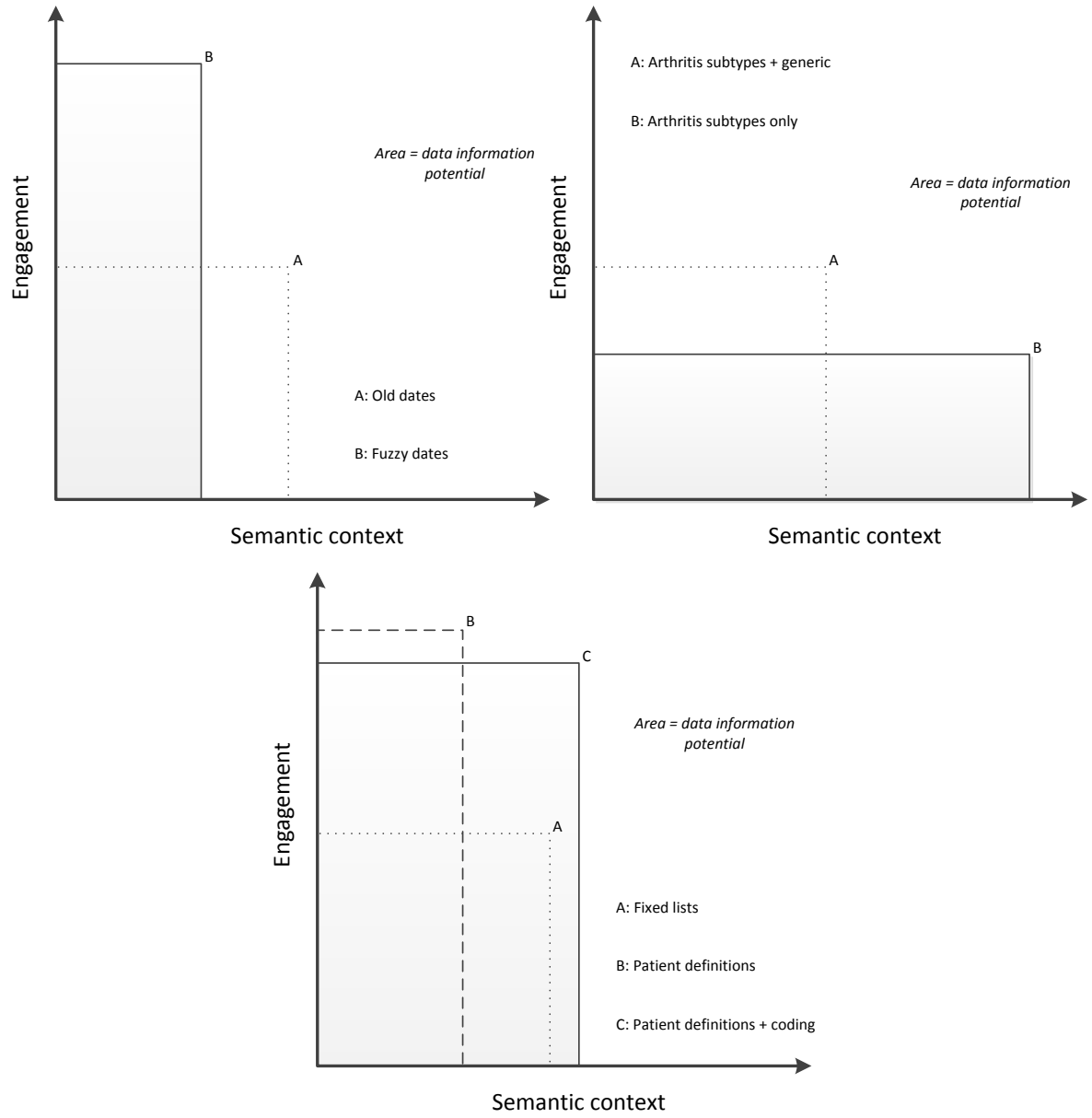



***Figure 3: Shifts in information potential, in the examples of fuzzy dates, generic arthritis, and patient-generated symptom definitions***

## *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 organisation 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 initial remark, it appears that *PatientsLikeMe* should be contrasted to other social media- and research-based organisations 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 organisation. The only source that the organisation has to find out about the patients – here collaborators upon which the organising depends (see also Kallinikos and Tempini, 2014) – 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 organisations 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 organisations but also overlaps with the ostensible development of "Big Data". A distinctive feature of these innovative data-based, or data-intensive, organisational 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 organisation 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 organisation 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 organisation 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 organisation 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, 2014). Social denomination is foundational for the form of '*computed sociality*' (Kallinikos and Tempini, 2014) 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,[47] which shows how organisation 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 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

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

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 here before opening a topic that clearly is 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 organisation 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.

# Patient Data as Medical Facts: Social Media Practices as a Foundation for Medical Knowledge Creation

**Jannis Kallinikos and Niccolò Tempini, LSE**

**Abstract**

This paper investigates a web-based, medical research network that relies on patient self-reporting to collect and analyze data on the health status of patients, mostly suffering from severe conditions. The network organizes patient participation in ways that break with the strong expert culture of medical research. Patient data entry is largely unsupervised. It relies on a data architecture that encodes medical knowledge and medical categories, yet remains open to capturing details of patient life that have as a rule remained outside the purview of medical research. The network thus casts the pursuit of medical knowledge in a web-based context, marked by the pivotal importance of patient experience captured in the form of patient data. The originality of the network owes much to the innovative amalgamation of networking and computational functionalities built into a potent social media platform. The arrangements the network epitomizes could be seen as a harbinger of new models of organising medical knowledge creation and medical work in the digital age, and a complement or alternative to established models of medical research.

## Introduction

In a seminal article, Susan Leigh Star showed how we might uncover the means and processes by which a scientific fact 'emerges which is simultaneously stripped of its

complexities and isolated from its relationship to a larger work/historical context' (Star, 1983: 224-225). In the wake of the so-called Internet revolution, with many organisations experimenting with unconventional approaches to knowledge making beyond the traditional boundaries of research and commercial institutions (e.g. citizen science, peer-to-peer production, crowdsourcing, social media), social scientists must renew this commitment. There is a need to capture and document what, in such contexts, would otherwise remain invisible or untold, in new web-based approaches to science making.

Some argue that medicine is about to be revolutionized by new technological capabilities that allow new ways of conducting research, and providing therapy and care (e.g. Topol, 2012). With this paper, we present a study of an organisation that draws on the social networking platform it has developed to pursue medical research that relies on data collected from a distributed, open, user base through patient self-reporting. At one end of this research process, there stand, as a kind of raw material, a myriad of patient observations about their life experiences. The final product at the other end is a number of peer-reviewed articles and other scientific publications. The outcome seems striking. Producing medical knowledge through the routine online involvement of patients provides a stark contrast to the complex, expert-dominated, prestige-laden, and costly institutional arrangements characteristic of medical research. It is thus reasonable to wonder: *How does this process actually happen? How can unconventional, Internet-based organisational forms address traditional expert problems (medical research) through the systematic involvement of non-professionals (patients)?*

At the least, we are aiming to explain the case in such a way that will address the following three interrelated concerns. First, we would like to know the conditions under which "non-experts" are involved in expert work. *How is patient participation organized and governed, to provide information on a reliable and continuous basis such that it can be used as the raw data for medical research?* Second, we search for the technological underpinnings of such an enterprise. *How are data collected and aggregated so as to document and analyze patient experience? How are social media and information technologies shaping human communication in this context?* Third, we seek to identify the broader implications. *Are traditional medical research practice and institutions going to be transformed by emerging research practices and organisational forms, and if so how?* This paper seeks to address these fundamental issues.

The central node of the network we focus on is *PatientsLikeMe*, a company that runs the key operations underlying the network.[49] Data collection relies on electronic questionnaires and forms that are made available online to network members. As data are collected, they are immediately and automatically aggregated and analyzed, on a continuous basis. In essential respects, the network epitomizes what the current literature (e.g. boyd and Ellison, 2008; Gerlitz and Helmond, 2013) construes as social media or social networking platforms. Patients are encouraged to enter data about their health status on a regular basis. The data thus made available are used for understanding and describing the patient experience at aggregate levels, with the aim of

[49] *PatientsLikeMe* is a for-profit company based in Cambridge, Massachusetts, USA. It was founded in 2004 and connects more than 250,000 patients (accessed July 28th, 2014). Further information is available at www.patientslikeme.com/about/

documenting the effects of medication, illness progression (or remission), and other medical conditions or relations of interest.

*PatientsLikeMe* uses the data thus collected for research purposes. To date, members of the staff have published 37 outputs – peer-reviewed articles in established journals, reports, editorials, and other formats. From the data collected from patient contributions, the research staff has been able to research a range of subjects. To name a few examples: symptom distribution discoveries (Turner et al., 2010; Wicks, 2007); omissions in patient education by medical practitioners (Wicks and Frost, 2008); distribution of social issues (compulsive gambling) across patient populations, and association to drugs (Wicks and Macphee, 2009); drug efficacy discovery through virtual clinical trials (Wicks et al., 2011). As a pledge to the patient communities it engages, the organisation makes most of the research publicly available (open access). In addition, the web-based system generates a wealth of information based on the patient-reported data and feeds it back to the community in the form of a large range of report pages, each dedicated to specific medical entities represented in the system (conditions, symptoms, treatments). No money is exchanged between the patients and the organisation. Patient participation in the network is voluntary, motivated by whatever rewards (cure, socialization, ailment knowledge, recommendations) patients can hope to obtain with respect to the serious conditions they are living with.

The rest of the paper is structured as follows. In the next section, we describe the standard practices of medical data collection and their institutional settings, and contrast them with the data collection arrangements of the network we study here. This is followed by an account of the research strategy, and the data collection and

interpretation methods. We subsequently provide a general overview of the network and its features. We then narrow down our focus to the details of one exemplary process of health reporting, that of symptom data collection. We describe how symptom-reporting processes take shape in the organisation, paying attention to how technological resources are leveraged to reframe the standard research practices of symptom recognition and recording. Following this empirical case, we discuss our findings in the light of the three fundamental questions we raised above. In doing so, we place our findings within a broader framework that links this case to some fundamental issues of technological and institutional change.

## Data and Data Collection Practices in Medical Research

*PatientsLikeMe* offers research services based on aggregating, packaging, and analyzing patient self-reported data. The organisation has been able to use its underpinning technological infrastructure to construct a unique offer in terms of the scale, longitude, and real-world reference of its data sets. The novelty and uniqueness of the network emerges against the background of the traditional conditions of medical research that we will now briefly characterize.

Medical data management has a long and complex history of non-medical expert involvement. Similarly perhaps to many other fields (e.g. Yates, 1989), structured health data collection has, over the last century, become a progressively more complex enterprise that has involved specialists other than doctors or nurses. It has had to take place in specific institutions. Only the realisation of hospital services in large scale has enabled the development and systematization of clinical statistics (Shryock, 1961). At the same time, stenographers, data editors, and data librarians have all played an

increasingly important role in the standardization, circulation, and storage of medical data in hospitals and other medical care settings (Berg, 1997; Bowker and Star, 1999; Timmermans and Berg, 2003). Medical data management specialists have helped systematize the recording, storage, and availability of data produced by medical experts, and have improved the comparability of records across units and contexts, an essential requirement for medical research (Timmermans et al., 1998). Yet, these specialists have mostly not been directly involved with the generation of medical data, which has remained a prerogative of medical experts and, crucially, the ineluctable outcome of expert knowledge application and expert judgment (Conrad, 2005; Dodier, 1998; Timmermans and Oh, 2010).

In the empirical part of this paper, we focus on symptoms data collection as the primary object of analysis, and it is worth briefly referring here to the differences between symptom detection in this network versus that in standard research settings. Traditionally, symptoms have to be discussed, assessed, and filtered through a clinical interview that takes place where and when the clinicians operate. In most situations, loci are traditional research and health care institutions (research hospitals, laboratories, etc.), and time is limited to the availability of the clinical professionals. Even when data collection concerns physical biomarkers and is automated through machinery, an operator needs to be available to operate the machine at the end of the data collection exercise.

In either a case study or a randomized control trial (RCT),[50] the patient shares and discusses the situation *in situ* with a clinician (nurse or physician). Only through this negotiation can a symptom become a legitimate, recognized fact. A symptom officially enters an information system as data only by the hand of an expert. By controlling data entry, clinicians have the ultimate word on what a symptom really is. The patient plays a dependent role in data collection, and only so far as perceptions and feelings are part of the phenomena under investigation, such as when reporting symptoms. The patient is otherwise excluded from the assessment of all other reportable and observable medical entities (clinical signs) and has no relevant role to play in measurement, nor in inference. The investigation of biomarkers and other observable clinical signs is performed by the clinician and their entourage of tools, the machines of the profession, through the full epistemological authority the clinician commands. This clinical, as we may call it, apparatus (Agamben, 2009) largely operates as an engulfing epistemological regime. It defines and interprets the evidence and prescribes the strategy and the objects of the clinical investigation. In the context of limited and fragmented clinical encounters, reductions in scope are necessary so as to obtain consistency and economy of efforts. Thus, the attribution of a local clinical

[50] The RCT is upheld by the evidence-based medicine (EBM) movement as the gold standard for medical research. The strengths of this quantitative experimental design for measuring the effects of a treatment derive from a number of features that aim to neutralize possible sources of bias. These broadly include the random assignment of study subjects to different groups, and the designation of each group to either the testing of a treatment or not. By comparing the results of the treated groups against the non-treated (control) groups, a hypothesis can be validated. Moreover, stable processing of experimental protocols can be protected by 'blinding', i.e. not disclosing particular information. Study subjects can be blinded (not told) as to whether they are receiving treatment or not (checking for placebo effects). Similarly, caregivers and researchers can be blinded as to who is administered what. Once blinded, actors are likely to refrain from altering protocols in order to obtain the favored outcome. For further elaboration, see the exhaustive guidelines by the group for the consolidation of trial standards, CONSORT (Moher *et al.*, 2010; Schulz *et al.*, 2010).

situation to an illness profile, medical category, or classification system is established by the clinician, and their expert knowledge. Against this backdrop, the routine generation of medical data by patients themselves represents an entirely new development that breaks with the history of medical records being used for research purposes and the institutional settings within which these records have commonly been generated.

***Patient-Network Data Collection***

*PatientsLikeMe* has developed a social media platform that a patient can join free of subscription fees. As the name suggests, the platform offers the opportunity to socialize with other patients going through similar life experiences. A patient manages a profile provided with common social media features: private messaging, broadcasting, and commenting features in addition to the self-representation tools of a profile picture, username and '*About me*' textbox. In addition, the system provides the patient with a set of health-tracking tools, whereby she can capture several aspects of her own health status. Examples of tracked aspects include the symptoms she is suffering from and their severity, the treatments being taken and related dosages or frequency, weight, labs and tests, and so on, along with many other health-related aspects. Patient members or their caregivers participate in the network voluntarily and generate data that are shared with the network.

The network that *PatientsLikeMe* has built contrasts with canonical models of medical research data collection (e.g. Berg, 1997; Marks, 1997; Timmermans and Berg, 2003) in a number of ways. *First,* the network breaks away from standard methods of generating medical facts, such as clinical interviews and RCTs, and the institutional environment of a hospital or other health care unit in which medical facts are

commonly embedded. The online platform represents a straightforward arrangement with rather few and simple patient network participation rules. The collected data are all generated through distributed input by the patients, from locations of their choice, and commonly from home. Through these arrangements, the network trespasses on the rigid boundaries separating medical expert practice and research – traditional loci being hospitals, primary care, and laboratories (Shryock, 1961; Star, 1986) – from the contexts of everyday living in which illnesses commonly manifest and patient experiences are lived.

*Second*, the network puts patients at the center of the task of data generation. In so doing, it violates or, at any rate, tweaks one of the pervasive customs of medical research, whereby data entry has been the exclusive prerogative of experts (medical doctors and nurses) as the ineluctable outcome of expert judgment. In several instances, the data collection features patient-generated health definitions. Original patient observations are assessed, further pursued, refined, and tested through in-house specialist-patient online interactions, before being incorporated into the system routines for further data aggregation. Still, most of the system routines related to data collection and analysis occur without the routine and direct involvement of clinical professionals. This is a clear departure from traditional medical data management practices in which clinicians are in control of data entry and clinical assessment while patients are relegated to a marginal and dependent position.

*Third,* data collection in the network is predicated on an inclusive, holistic understanding of health that goes far beyond the medically recognizable conditions of particular diseases. Data collection is, to a degree, use-agnostic, open to the recording of

rather broad aspects of patient life. In the hope that all data might turn out relevant, the network seeks to capture a wide range of circumstances, beyond those that medical researchers would traditionally earmark for data collection in the context of specific research undertakings. As we show in the context of symptoms data collection, patients can choose to track a range of symptoms that is much more granular and extensive than expert terminologies often allow for.

*Fourth*, data collection is longitudinal, encouraging reporting at all stages of patient life. It is also continuous, seeking to obtain patient inputs as frequently and regularly as possible. The longitudinal and continuous data collection is based on the assumption or belief that it is worth capturing the patient experience in significant detail, transcending the standard focus of most institutional care and research. Patients are free to enter data as often as they believe necessary, as the technology automates many of the transactions involved.

Taken together, these attributes describe a new and different way of organising data collection for medical research. They lie at the heart of the network and the value it generates for several network stakeholder groups, including the company owners and employees, patients, medical research communities, and pharmaceutical companies. Little wonder that such attributes have been variously anticipated by the contemporary medical research and care practice. Giving patients greater leeway in diagnosis, therapy, and even disease management, observing the progression of diseases and patients over longer time scales, and integrating facts about life and disease have all been developments, in varying degrees, of current practice (Berg, 2004; Clarke et al., 2003; Conrad, 2005; Timmermans and Berg 2003). Similar views have been characteristic of

the wider political discourses in which health care has been embedded over the last few decades (Hasselbladh and Bejerot, 2007; Tousijn, 2002). In this respect, the network we study both reflects and embodies wider assumptions that are diffused throughout current practice but also society at large. Yet, through the coordinating framework of social media, these distinctive attributes have been catalyzed in new and interesting ways (Prainsack, 2014). The network exemplifies a new architecture for organising data collection, and new capabilities for analyzing and assembling evidence that require in-depth investigation (Star, 1986). As we hope to demonstrate throughout this article, the distinct configuration of the network we study here derives from the flexible forms of interaction enabled by social media and the innovative deployment of the functionalities afforded by current computing and communications technologies (Jonsson et al., 2009).

## Research Design and Methodology

A participant observation case study was conducted between September 2011 and April 2012, over 26 weeks, at the headquarters of the organisation. One of us participated in work activities, mainly as an R&D team member. He was involved in several projects, while at the same time allowed to exercise great discretion over the time and resources committed to each project. Participation took the form of regular office hours, five days a week, and occasionally entailed acting as a delegate, representing the organisation at conferences and in meetings or calls with external guests or partners. The researcher had access to resources that a regular research team member would have.

Such an intensive involvement in the organisation allowed the researcher to join forces with most of the employees working at the company's headquarters (around 30-40 members during the period of observation). Beyond the informal observation of work and conversations, the researcher participated in numerous formal meetings – 128 in total – of different kinds, from project-specific task force meetings to stand-up developer meetings, release demo meetings, and company meetings. In addition, he was able to collect data from documents, screen snapshots of user- and admin-facing systems, slide-show presentations, internal e-mail messages and conversations, and the work that the researcher himself produced for the organisation. During his time on site, the researcher logged his observations, in the form of notes typed on a laptop using dedicated note-taking software. This software log was constantly at hand for recording immediate observations and reflections. Even during regular working hours, the researcher was relatively free to detach himself from the regular workflow, to develop notes and reflections that he felt needed prompt recording and elaboration. Additional reflections were logged off site – at evenings and weekends. Tentative interpretations of what he felt were compelling observations and events in need of further explanation were developed *in situ*, crosschecked, and stored (Aaltonen and Tempini, 2014; Sayer, 2000; Van Maanen, 1979, 1993).

Due to the size of the workforce, most of the employees of the company, at all levels, were interviewed, some twice, based on their perceived proximity to the issues under research, and institutional knowledge and memory.[51] Interviews were semi-

[51] The researcher held 30 individual interviews, with an average duration of 60 minutes. Snapshots and written notes added up to 665 analytical episodes stored in the electronic log.

structured, yet the interview guide was prepared anew for each interviewee to accommodate their role and work, and the progression of the empirical study and collection of facts to that point. Interviews were held throughout the empirical study, but with more than half of them concentrated over the last month. Following the fieldwork, most of these interviews were transcribed and analyzed together with other written and documentary evidence.

With participant observation being the key vehicle of data collection, this should indicate that the empirical investigation featured an exploratory case study research design (Yin, 2009). The state of the field on such novel developments did not provide us with firm theoretical propositions with which to link our data collection (Yin, 2009). Embedded, observational case studies are an adequate research approach for developing new explanations in the absence of a theoretical framework that stipulates the conditions for research (Sayer, 2000).

In the context of medical research carried out through social media and patient involvement, our first immediate goal was to assemble empirical observations with the view of addressing the questions we raised in the introduction. The ways these questions were framed (see our introduction above) directed our attention to the means, processes, and techniques by which the company and the network organized, fragmented, and distributed its data collection work, and its data processing and aggregation. Intermediating social interaction through text, measurements, categories, and classifications, the network had to be studied by putting the processes of the construction of health descriptions and symptom detection at the center. The stage of data collection followed by and large what, in current grounded theory jargon, is

referred to as theoretical sampling (Corbin and Strauss, 2008): the period of participant observation entailed a steady calibration of data collection with emerging interpretations. Our data analysis and interpretation continued after the fieldwork period, mainly through the crosschecking of the empirical material with a view to identifying a consistent narrative about the phenomena under investigation. In this process, we compared our empirical findings on the processes of data collection and analysis used at the field-site to data collection processes depicted in the literature on medical research and medical knowledge creation. Much of that comparison took place against a wider understanding of the role of social media, data, and computation. After several iterative readings and analyses, we selected the most relevant pieces of evidence and assembled them into a case study narrative, following retroductive logic to produce our explanations (Sayer, 2000).

In such a unique and innovative case as *PatientsLikeMe*, it was clear from the beginning that many different questions could and should be asked. The network presents itself as a disruptive and unique organisation at the crossroads of patient advocacy, evidence-based activism, health care provision, and the pharmaceutical industry. In this purview, it is compelling to prefigure issues of democracy and representation, for instance, against which much of the literature has contrasted similar organisations and initiatives (e.g. Epstein, 2008; Rabeharisoa et al., 2013). However, it became clear to us that none of the central issues with which we were concerned could be satisfactorily pursued in the field without first accounting for the premises of systematic patient involvement in medical knowledge creation, and the role technology plays in this process, both as a platform of sociality and as a computational force supporting data collection and, critically, data aggregation and analysis. Both research

interests (sociality and computation) shaped our interpretations of the documents we collected, the viewpoints we recorded in the interviews, and the explanations we advance in this paper.

## Empirical Findings

Self-tracking can be useful to patients, not only for health monitoring, but also for socialization opportunities (Treem and Leonardi, 2012). New lab results, disease courses, or other unfortunate health developments can be important subjects for interaction with other patients. For many patients, *PatientsLikeMe* is primarily a network for support, solidarity, empathy, and companionship. A patient can make use of a number of filters, provided by the system, to browse the network member base and find other patients confronting similar health situations. The efficiency of the system in connecting a patient to other patients with whom they share relevant characteristics (e.g. condition, co-morbidities, treatment regimes) very much depends on the amount of data that the patient inputs into the system. The more data a patient enters about her own situation, the more the system is able to draw connections across the member base.

For *PatientsLikeMe*, health self-tracking represents the possibility to collect valuable views on patients' health status. A host of tracking tools is at the patient's disposal. The patient can enter data autonomously and continuously, generating data over time – traditionally a very expensive and difficult-to-accomplish feat. This can happen whenever the patient finds it most feasible or useful, according to her own routine. System features do encourage data input at regular intervals through user

interface (UI) notifications,[52] but the network aims nonetheless at maximizing data collection opportunities. Depending on their condition, patients may lack the time, energy, or even the opportunity to enter data at consistent intervals and volumes. The system therefore allows data inputting at irregular intervals, privileging input volume over timeliness.

Patients can explore information about their own health through various forms of data output. Through data aggregation techniques, the system dynamically constructs and displays profile pages on specific kinds of medical entities (conditions, symptoms, treatments, labs, and others), represented in the form of scores, descriptive statistics, and visualizations. Patients can thus browse a range of reports that put their profile data in perspective and offer a complex picture of the individual. Patients can also browse data representing the health aspects of entire patient populations. Patients access a wealth of information that the system generates by aggregating the data contributed by patient members across the network. Browsing a complex and dynamic network of links, a patient can quickly navigate from her own individual profile to population-level 'symptom (or condition, or treatment) report' pages. A 'symptom report page' shows, for instance, descriptive statistics such as the distribution of symptom severities (number of patients reporting severe, moderate, mild or no effect), the demographics of the affected population, and a list of the most popular treatments that patients associate with the symptom. It also shows links to the profiles of other

[52] Only under particular conditions, such as for instance when a patient has not updated her symptom data for more than 30 days, does the system activate constraints on the UI that limit functionality. In the context of symptom data collection the system will require patients to update their symptom data before performing other operations like tracking new symptoms.

patients suffering from that specific symptom. Because of this webpage structure, the platform provides the patient with information that can help her to understand her health situation, and links promoting and aiding social interactions with others who are similar.

The system generates these up-to-date statistics dynamically.[53] The patient can thus explore parts of the organisation's database, in 'sliced and diced' form, by navigating a web of interlinked pages. Patients should then be able to access information that could help her to make sense of specific health situations. A patient can add specific items (e.g. a symptom) that she wants to track on her profile by following links in the item's report page. In so doing, the patient tailors the system to track all the aspects that she deems relevant to her life experience.

Enabling patients to track all aspects that they judge relevant is a strategic goal for *PatientsLikeMe*. The potential for clinical discovery – for collecting the *'gems out there'*, as one top executive defined the rare or insightful correlations or events the company hopes to discover – makes the case for this ambitious distributed data collection architecture. An underlying assumption is the idea that, in respect to a given medical issue, there can be revelatory cases out in the world, and these cases can be documented if the appropriate communications infrastructure is developed. In order for these cases to be discovered, however, the system will be more effective if its data

[53] When patients navigate through pages such as symptom report pages, report data are aggregated 'live' through database queries triggered by the execution of the web-application code. The aggregated data are then stored in the cache for as long as they are still up to date, to improve performance (i.e. lower page load time). Depending on the kind of data, they can be cached for between three hours and a month.

collection remains open and, at the same time, sensitive to a wide range of phenomena, avoiding the over-fitting of events into pre-existing categories. This goal of distributed research data collection adds both promise and burden to organising collection by means of systematic patient input.

The data collection architecture has therefore been developed with the goal of detecting and mapping clinical diversity in patient experience. Important clinical events could easily escape being recorded. Often, they do not manifest evenly in a patient's life. Also, they may be thought of as singular or irrelevant, and forgotten by the time of the next data collection opportunity. Even small symptom signs, which may seem *prima facie* irrelevant, may have significance and eventually amount to a premonition of future developments (Tempini, in press). The open and distributed data collection architecture makes it possible for phenomena to be documented that usually escape recording in traditional settings because they may seem irrelevant, ephemeral or are not easily mapped onto medical experts' categories and classifications. An open architecture can also empower patients. One top executive commented that the system has a fundamental capability to record the patient's voice, with its concerns and insights, in the form of data entries: *'That data is a rich field of information to look at and understand patient concerns'.* The data that patients input into the system can represent needs and concerns that, in the past, were left unvoiced: *'Some of the stuff is not necessarily categorized today in medicine'*. As we show in detail in the next section, the system architecture allows patients to create new symptom categories and to aggregate data inputs into new categories, affording a bottom-up development of a medical categorization system.

Obviously, the potential for open and sensitive data collection is difficult to realize. As the aforementioned top executive said in an interview, '*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) or really interesting stuff'.* Successful data collection requires not only that the system adapts to the life contexts of patients, but also that the researchers devise strategies for reducing biases, errors, and conflicting interests. This is the concern of the complex processes of category review and validation we discuss in the next section.

***Symptom Data Input***

Each kind of medical entity (symptoms, conditions, treatments, etc.) is described by some defining characteristics. These inform the development of the data collection system. Symptoms are ontologically simpler than other medical entities and for that reason suit the purpose of the present paper. Indeed, conditions have more cumbersome and ambiguous ontological histories (Bowker and Star, 1999), while treatments require more complex data models specifying many parameters (dosage, form, frequency, etc). In the context of symptom tracking, the *PatientsLikeMe* system allows the patient to input severity ratings (none, mild, moderate, severe) and add treatments with which the patient associates the symptom. Figure 4 depicts the possible associations that patients can draw between a symptom and a treatment.

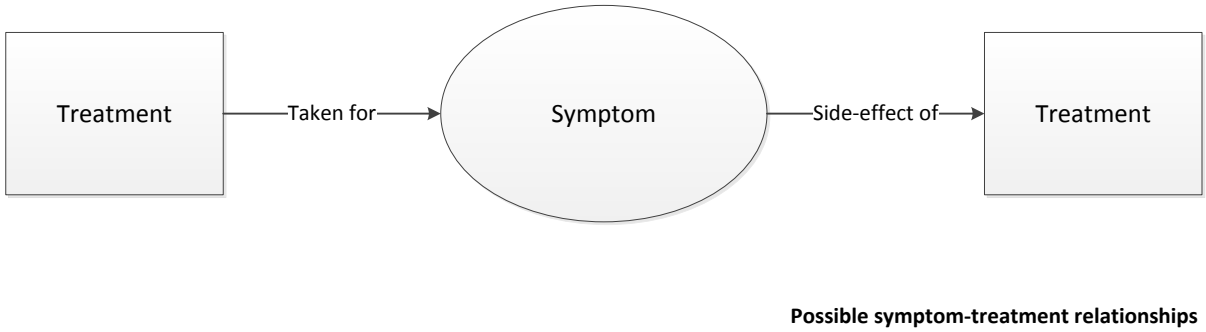



***Figure 4 – Two possible ways of drawing a symptom-treatment association***

A symptom can be associated with a patient profile in three ways. In the first two ways, the system automatically assigns symptoms to a patient profile. First, there are general symptoms. These are symptoms that are expected to cut across the spectrum of all patient experiences and are assigned to patients of all conditions: *anxious mood, depressed mood, fatigue, insomnia,* or *pain*. The patient is encouraged to track these generic symptoms, because they constitute a common denominator of basic patient life experience. Second, another set of symptoms are automatically added by the system to a patient's tracked symptoms. These are condition-specific symptoms and depend on the conditions that a patient adds to her profile.

Conditions represented in the system are administered through configuration files. The configuration file holds the 'genetic code' of a condition: it stores a number of relevant pieces of information that trigger a number of links or features across the system. Among other things, the staff can store in the configuration file a list of condition-specific symptoms – these are symptoms deemed to characterize the common experience of patients suffering from that condition. As an informant explained, in this way the system is able to automatically adapt and *'serve up'* symptoms to patients: *'The*

*only way we have to serve symptoms up for patients in relationship to a condition is to identify them on the admin tool as the primary symptom'.*

When a patient then adds a condition to her profile, the system assigns the set of condition-specific symptoms to the symptoms to be tracked. If a patient adds a condition that does not have condition-specific symptoms stored in the configuration file, the system refrains from assigning additional symptoms to the patient profile. The identification of the symptoms that are specific to a condition is a labor-intensive task requiring a considerable amount of research. Only a fraction of the conditions stored in the system have so far been assigned condition-specific symptoms. This usually occurs through funded projects that allow the staff to undertake the required research. The list of condition-specific symptoms is compiled from various sources that describe the common experience of a specific condition (more on this later). As a member of the integrity team explained,

> *We are trying to pull those symptoms from an architecture of reference in science; it's sort of saying "ok, what are the ones* [symptoms] *that most commonly people might have experienced".*

The third way in which a symptom can be associated to a patient profile is by the patient herself, adding symptoms to her profile through a link in the symptom report page (the page dedicated to the dynamic description of a symptom). Symptom report pages can be found through a search feature, by which the patient can find out whether the symptom is already present in the database and, by accessing its report page, see how other patients experience it. By adding the symptom to the profile of the patient, the system enables the patient's experience to be linked to an already existing symptom

category. In this way, it is possible for the data collection to aggregate data consistently. The experiences of different patients are thus made similar and comparable through the mediation of the system. The structured nature of the data – with labels and other data fields – makes it possible to aggregate and compare the data that one patient enters with that entered by other patients in the network.

The system uses a number of techniques to help the patient match their symptom to one recorded in the database. As the patient searches for a symptom in the search box, typing the search query letter by letter, a drop-down list starts to show dynamically parsed, instant results.[54] The tool, powered by spelling-correction features, highlights the matching words in the instant results.[55] Clicking one of these results takes the patient to the symptom report page. On that page, she can review the information the system displays about that symptom, consisting of the following elements: first, a symptom description, presented in a verbal, free-text form; second, the distribution of symptom severities on the NMMS scale (none, mild, moderate, severe) across the member population; third, the distribution of treatments associated with the symptom by other patients across the member population; fourth, links to a few profiles of patients experiencing the same symptom and a link to the complete list of all symptom-

[54] This feature is similar to the instant results outputted in Facebook's or Google's drop-down search menus.

[55] For instance, if one types the wrongly spelled 'ancious' in the search box, the drop-down menu offers the following results with associated patient populations. The highlighting shows the matching element:

*Anxious mood* *251331*
*Less anxious* *4*
*Stiffness in legs when anxious* *2*
*Anxious mood in the morning* *2*
*ancious isn't in our system. Submit a request to add it*

related patient profiles; fifth, links to a few forum posts related to the symptom and a link to all symptom-related forum posts.

If the patient is not successful in matching her individual case to an existing symptom, the system allows her to initiate the creation of a new symptom. The new symptom first undergoes a review by *PatientsLikeMe* staff. After the review process, a new symptom record is created and fed back into the system. A symptom report page is automatically generated, and other patients will then be able to add this symptom to their tracked symptoms list. In Figure 5, we depict the different mechanisms by which a patient profile's list of tracked symptoms is completed.

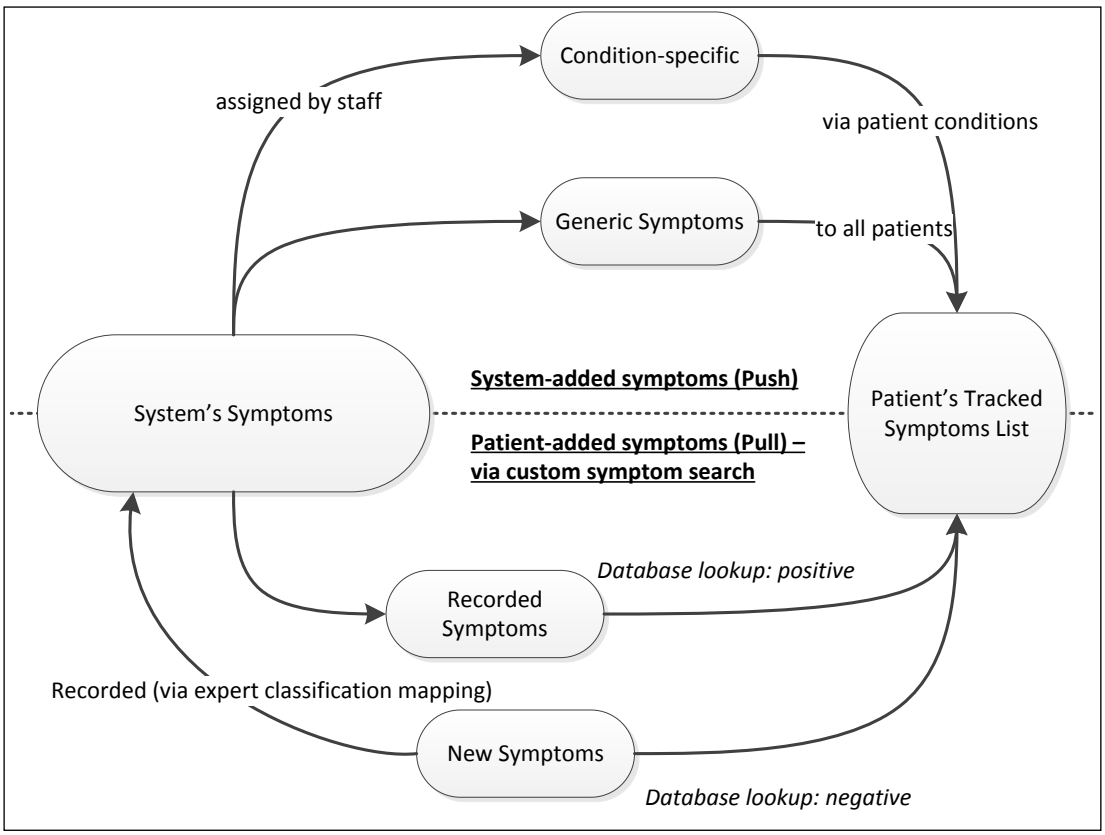



***Figure 5 – Mechanisms for adding a symptom to a patient profile***

As referred to above, if the patient is unable to find a symptom matching her experience, a link in the drop-down menu of the search box allows her to request the creation of a new symptom and to provide its definition: '[symptom] *isn't in our system. Submit a request to add it.'* Upon submission of the request, the patient is informed that the new symptom is pending review from the staff.[56] At the other end of the system, the new symptom shows up in a dashboard used by the *PatientsLikeMe* staff in charge of the ongoing curation of the medical database.

On this dashboard, staff review new symptom requests as well as other new item requests (treatments, conditions, hospitalizations, etc.). The open, distributed, and bottom-up categorization strategy is applied to all kinds of medical entities in the system. When a staff member (most often a registered nurse, pharmacist, or biologist) reviews a case, she can follow a number of alternative courses of action. Often, the definition a patient provides of a symptom is not self-explanatory. Patients might describe something using unclear wording. Sometimes, they may propose as symptoms things or events that are not symptoms. It then becomes necessary for the staff member to reconstruct the context of the patient's request. The staff member can send a private message to the patient asking for clarification or more detail. Through a number of messages, the staff member performs a short, mediated interview seeking relevant evidence so that she can decide how to manage the request. As a nurse and clinical informatics staff member explained,

---

[56] The message reads: *'You have successfully added [X]. A patientslikeme staff member will soon review your addition and add it to our global symptom list. You will receive a private message when this process has been completed, and you will be able to add it as a symptom.'*

> *I have to iterate with them: "*[…] *you know, based on what you're showing me and what your picture is, this is what I think it* [the new item] *might be but I could be totally wrong; just let me know".* […] *I'm guessing what's happening based on my nursing background, and helping them to paint a clearer picture for everybody else.* […] *There are certain pieces that they* [the patients] *don't necessarily think are important to add to their profile that are helpful for other people if they know the whole story.*

The staff member also looks up the requesting patient's profile, in order to find clues that could explain what the patient is experiencing and trying to communicate. The staff member embarks on an investigative task, drawing possible connections between the conditions a patient is suffering from, the treatments the patient is taking, the surgeries undertaken, the number, sequence, and dates of diagnoses received, and other relevant information the patient has spontaneously stored in the '*About me*' textbox. Useful context can be provided just by some biographical information – '*there's just something about knowing about their age, about the other conditions they have'* – and health history:

> *Symptoms may be different because of their* [patients'] *condition. So, if someone puts in something vague that might be condition-related and I can check the profile, and I see they've got this condition, it means she is probably talking about the symptom in this context.*

To complement the information about the patient context sourced from conversations and the patient profile, the staff member consults external resources that

can range from PubMed and other E-Medicine portals to Wikipedia, UMLS meta-thesaurus and results from Google searches. Through this process, the staff member seeks to progressively define the nature of the item that she is negotiating about with the patient. The ontological status of medical entities themselves – conditions, syndromes, symptoms – is often ambiguous and disputed: *'There is one thing that we are always parsing around here. What's a symptom and what's a condition… Sometimes the patients do not necessarily make the distinction.'* Sometimes the staff cannot clarify the case and it becomes apparent that a prompt decision will not be reached any time soon, such as when a patient simply does not reply to a staff member's questions. In this case, the staff member *archives* the item into a dedicated folder for eventual future follow-up.

When, instead, a decision is reached about a symptom request, the staff member takes one of a number of different actions in the dashboard. One is to *merge* the new symptom into an already existing one. It can happen that a patient fails to notice that the symptom already exists in the database. Mishandling the search feature through a major misspelling error or an incomplete definition may lead a patient to submit a symptom request that can easily be solved. In this case, the staff member merges the new record to the original one: *'I know the context and I have a couple of different pieces of the equation that I might be able to say "yeah, ok, merge".'* The new label created by the patient upon submitting the symptom request is discarded, and the patient's data are aggregated with the data for the group of patients associated with the existing symptom. Often, merge actions are laborious, and involve the inspection of the patient profile or interaction via messaging features. The staff member continues searching to find out whether the patient experience corresponds to and can be assigned to a specific symptom. For instance, one staff member realised that 'swelling' in fact meant 'injection

site swelling' by looking at the patient profile and noticing that the patient's treatment entailed subcutaneous injection:

> *I could check their profile and I could say 'Oh, that person's on Copaxone'.* […] *So you can bring those patients together in those reports; so now these patients are grouped together; it's not just this person has got a side effect of swelling, it's injection site swelling; you get that context from the profile.*

A second course of action the staff member can take is to approve the request and *create* the new symptom. The staff member produces a short description of the symptom, based on the information the patient provided and what it was possible to obtain from other medical sources. The new symptom becomes part of the symptoms database, a symptom report page is automatically generated, and other patients will be able to search for and add the symptom to their own profiles.

Sometimes patients enter multiple symptom entities in the same text string.[57] In this case, the staff member *splits* the symptom into more than one symptom. If necessary, a new symptom is created, but in most cases splitting a new symptom item involves summoning existing symptoms. Through merging and splitting symptom requests, the patient profile can be redirected or subsumed under appropriate categories and thus become an object of aggregation. Tools for merging and splitting symptom requests were not part of the early features for administrating the

[57] For instance, one symptom for review could in fact be a string containing two symptoms, such as 'toothache cognitive impairment.'

*PatientsLikeMe* system. They were developed in order to streamline and automate some standard, common operations that previously depended on patient actions:

> *Let's say they accidentally entered a treatment as a symptom. There is no way for me to* [change it to a] *treatment from* [a symptom] *entry and I didn't want to code it up as a symptom and you can't delete it because it's patient data.* […] *I would have to message the patient. We then helped build tools like splitting and merging.* […] *We now have the ability to merge something. If someone puts in 'Fibromyalgia, head pain, headaches', now we can split it into these different categories of already existing databases and make new ones out of it too.*

As a course of action unfolds, the staff member keeps the patient informed and provides an explanation of the action taken. Patients often react if they believe the label they provided still best describes their experience, and it can happen that a staff member will make an incorrect guess. Keeping the patient informed on changes encourages feedback for the actions taken. In the following two diagrams we summarize the interactions we have just described. Figure 6 depicts the operations involved when a patient adds a symptom to her profile that is already present in the system. Figure 7 depicts the operations involved when a patient adds a new symptom to the system. We highlight as 'controlled computation' the steps of the routine that come under expert review. Through the reconstruction of these flows, the organisation has engineered a pattern of mediated and linked interactions that utilize advanced data representation techniques to support the patient in the process of data collection. In the cases where automated support breaks down, technology enables, as we have seen, the intervention of a clinical professional and the repairing of the process through several techniques of

disambiguation, including patient-staff remote interactions and the use of a range medical resources and data representations. Breakdowns can happen because the automation is not sufficient to help the patient find the appropriate category, or because the patient experience does not conform to other experiences captured in the database.

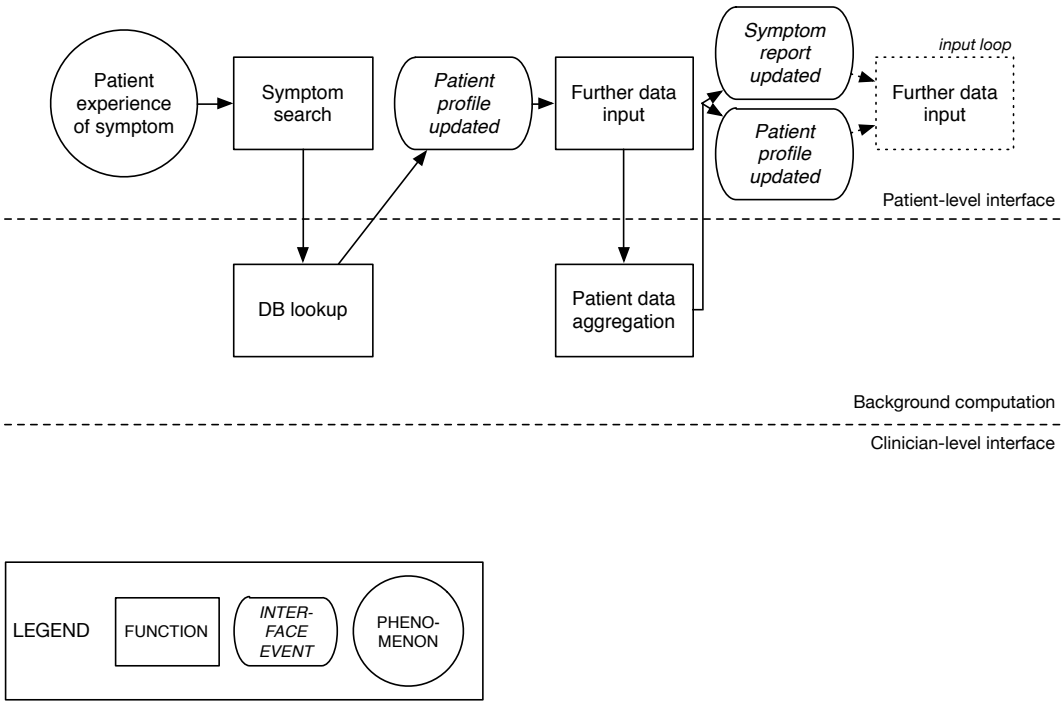



***Figure 6 – Operation of adding to a profile a symptom already present in the database, divided between patient-level interface, background computation, and clinician-level interface***

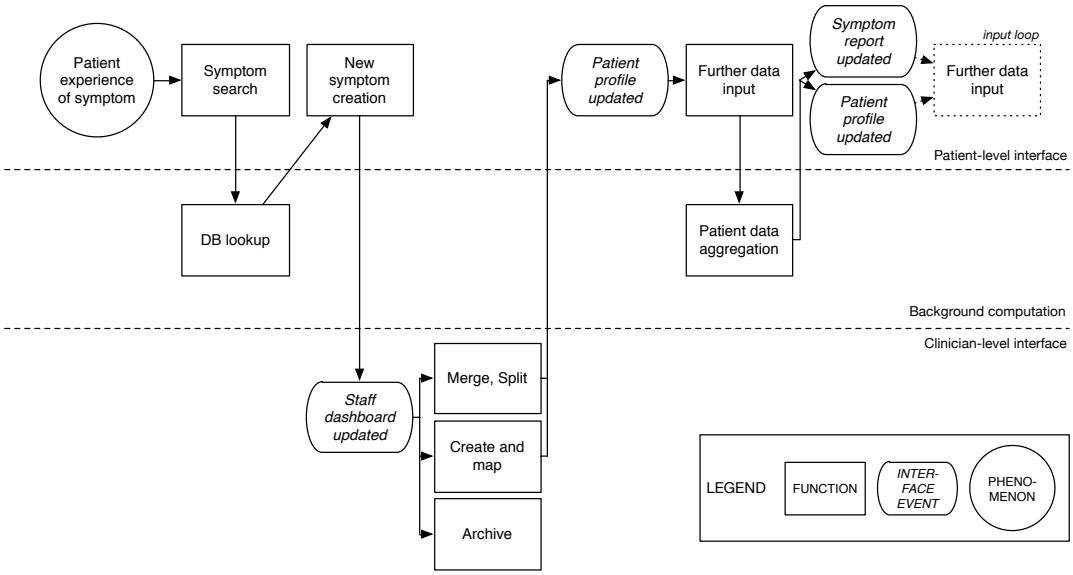



***Figure 7 – Operation of adding to a profile a symptom not already present in the database, divided between patient-level interface, background computation, and clinician-level interface***

To allow new phenomena (patient experience forms) to be detected and emerge through a bottom-up process is a Janus-faced accomplishment. Indeed, to be useful in medical research, new labels need to be made sense of – and meaning arises only if connected to medical knowledge. Therefore, on the one hand, the novel aspect (difference) of a phenomenon needs to be brought to the fore and highlighted, through dedicated definitions and data representations. On the other hand, it is important to place the phenomenon among what is known. New phenomena are only new to a limited extent: at a cost, much can be reduced and subjected to the existing ontology, if need be. Based on their interaction with the patient and apprehension of the illness details that make up the context of the patient's life, the staff member maps the new symptom record to a symptom represented in the expert classification systems (SNOMED, ICD10, ICF and Meddra LLT) through coding. This operation enables the dovetailing of the patient-generated definitions to established expert definitions, that is,

of the patient experience language to the clinical professional language.[58] Often patient-generated symptom definitions describe experiences with more nuances or detail than the definitions employed by expert classification systems. For many patients, some nuances are relevant that experts would not recognize as such. Preserving patient definitions means preserving information that can be meaningful not only to patients but also to researchers. As one informant explained, while the system allows the researcher to see the hidden associations between analogous symptoms, it also preserves the patient voice that could be a source of further differentiation: *'You get down to the one that the patient actually told us about in their own words.'* The coding enables researchers to aggregate different symptoms under a more generic, expert category, and combine the respective data as instances of the same phenomena.

This open, distributed data collection architecture has, over time, come to host about 7,000 patient symptom definitions. Many of these definitions differentiate and specify phenomena along more ordinary medical dimensions; in others, social, personal, and emotional meaning prevail, testing the boundaries of established medical concepts and categories. Collecting and storing perceptions and experiences for the most varied and often multiple reasons, patients overlay the traditional and restrictive condition-treatment-symptom architecture of the patient experience with spurious – but phenomenologically connected – phenomena of everyday living. Coded against the ICD10 code R45.3 'demoralization and apathy'[59] are patient concerns of various kinds: *'loss of ambition, loss of interest, life appeal, not caring further if I die, apathy,*

[58] For instance, the symptom 'anxiety with telephone' is mapped to 'specific (isolated) phobias'.
[59] http://www.icd10data.com/ICD10CM/Codes/R00-R99/R40-R46/R45-/R45.3

*environment, no motivation, inability to initiate tests, disorganized…'.* There, in this messy, laborious, and expensive data collection exercise stands the potential that the network is trying to cultivate, for grasping knowledge that lies at the boundaries of social and linguistic conventions, yet is linked to established medical definitions:

> *Certainly I think it's great that we have a less clinical database in here.* […] *Because that's what we are trying to do is use your voice, patient voice, patient-centered data, all these terms we use. It would be kind of hypocritical to create databases that only we decided what would be the entries in them.* […] *you code against happy and unhappiness or social behavior* [issues]. *That is not something that is going to be in any clinical book; it is not going to be ICD. It is not going to be like that but you code it with something similar so that it gets grouped with socialization disturbances or behavioral disturbances and social stuff, and it is all in there.*

## Discussion

The empirical evidence presented in the preceding pages describes the processes and arrangements based on which data on symptoms and patients are collected, ambiguities in the process of symptom mapping are negotiated or settled, and data are made sense of, at both the individual and aggregate levels. In what follows, we draw on this description to address the three fundamental questions we raised at the outset of our investigation, concerning (1) the premises of patient participation in the network, (2) the technological underpinning and organisational arrangements

underlying patient data collection, and (3) the putative implications these developments carry for medical practice and institutions.

***Network Patient Participation***

What seems to strongly differentiate *PatientsLikeMe* and the network it governs from the canonical models of medical research reviewed earlier in this paper are the largely unsupervised data entry by patient populations and the concomitant modest expert contribution that underlies the online process of symptom mapping. The unsupervised data entry by patients establishes the conditions for a diversified information inflow that captures facets of patient life that have hitherto remained beyond the scope of expert medical work and research. It is this objective of capturing the details of patient life and the events that punctuate their everyday *en masse* (to obtain the '*gems out there*') that pervades the network and lies at the heart of the distinctive contribution it is making to medical research (see also Tempini, in press).

The objective of capturing the patient everyday in these terms requires the steady and reliable procurement of patient data. Organising patient participation on this scale is a complex and delicate accomplishment. While massive and largely unguided, the data entry is nonetheless carefully crafted and architected. The mediation of patient life occurs via an elaborate grid of data fields and categories (e.g. generic and condition-specific symptoms) through which the system and the platform encode existing medical knowledge and other facts of patient life (e.g. biographies, treatments etc).

At the same time, the process of symptom mapping remains open to recording aspects of patient life that do not fit the prescribed categories of medical knowledge.

This is accomplished through patient-staff online interaction and a navigational structure through which the process of symptom mapping and creation is organized. Figures 3 and 4 illustrate the pattern of these interactions beyond established medical categories and the series of steps through which patients and staff members negotiate the reality of the patient experience. The objective of reaching beyond the boundaries of established knowledge is also assisted by the links between patients themselves. Through these links, patients can trace aspects of their patient life that might otherwise have escaped their own awareness or observation. The dual accommodation of the requirements of structured data input and the open character of the events that punctuate patient life is the distinguishing mark of the network.

All these vital operations are, in turn, critically dependent on the steady inflow of information, without which the entire system would collapse in one blow (like a spacecraft without fuel). Ensuring a steady level of patient contributions is a delicate task that is sustained, as we show below, by the inventive deployment of the social media platform on which the entire network relies. Web technologies make it technically possible to collect open and longitudinal data but how does this become practically and socially possible? By what means is patient activity in the network sustained? Patients contribute to the network voluntarily and for multiple and often unexpressed reasons, according to their life schedules and priorities, while many of them are dealing with the dramatic implications of their illnesses. Still, *PatientsLikeMe* depends on patient contributions, as it does not source health data by any other means. With no patients contributing their data over time, the organisation would collapse.

Elements supporting the steady inflow of data are, in the first instance, the very social features and interactions that the platform makes available. As indicated, patients enjoy a range of standard social media tools and features that facilitate communication with others in similar situations. However, more than the tools provided to the patients to sustain online conversation, what is critical is the way the platform supports their connection with other patients. In this sense, the platform is an environment that continuously generates possibilities for interaction (connections) and records their outcomes. Patients are linked to specific forum rooms according to the conditions they add to their profiles (they are free to participate in others too). Also, they search for other similar patients through the patient search feature. Patients can filter the user base according to health parameters. The feature is more effective when the patient has entered data about herself, as the system is then able to use those pieces of data to pre-select certain filters. Crucially, for a patient to find someone else, other patients must have entered data about themselves. Conversely, in order to be found by other patients, a patient must have entered data about herself.

Even more powerful than the patient search feature are the links to other patient profiles that pervade the platform on many of its pages, and by means of which data collection is strongly coupled to interaction possibilities. In our description of the symptom report page, we highlighted how the page embeds a host of links that allow a patient to navigate to other patient profiles. These links are as numerous as the number of patients taking a certain treatment, reporting a certain symptom severity, commenting about the symptom in the forums, and so on. The platform, through dynamically constructed pages and database associations, continues to reshape the linkages between one patient's experience and the experiences of other patients. The

range of links that reflect possible connections between patients, worked out on the basis of aggregated data operations, constructs a web of socialization possibilities that become a steady source of patient activity on the platform (see also Tempini in press).

### *Technological Underpinnings and Organisational Arrangements*

In some basic ways, the technological underpinnings of the network coincide with its social media platform, split into patient and clinical interfaces that are supported by a series of background database operations (see Figures 3 and 4). At first glance, one might, perhaps rightly, conclude that patient participation and the linking possibilities it affords depend on a set of straightforward networking options or capabilities typical of web technologies. Patients are put in touch with each other in various ways and explore their links to other patients themselves. The platform intermediates their exchanges.

In fact, much of the social media literature deals with these kinds of social links enabled by social platforms (boyd and Ellison, 2007; Gerlitz and Helmond, 2013; Morris, 2012). Studies have shown how social media enables certain interactions that assist with knowledge production and collaboration within, across, or beyond organisations (Faraj *et al.*, 2011; Majchrzak *et al.*, 2013; Treem and Leonardi, 2012). Actors in organisations use social media to reach out to heterogeneous, public tools that afford several kinds of associations. In this regard, social media platforms afford association of 'people to other people, people to content, or content to content', as Treem and Leonardi put it (2012: 162), to support social connections, provide access to information, or enable emergent connections through rankings and recommendations (see also Scott and Orlikowski, 2012).

Yet, our reflection on the connections produced by *PatientsLikeMe* that we provide above takes these insightful observations a step further. In our case, patient links are made possible and realised through a series of computational operations, whereby data associations and data manipulation become the principal means for constructing social linkages. The links that are drawn through scores and numbers in the symptom report pages are produced through the filtering, juxtaposing, and aggregating of specific patient data. It is through these data computations – of two or more data tokens belonging to different patient profiles – that a third entity is produced (scores, counts – e.g. Desrosières, 1999), whereby associations of one patient with the life paths and experiences of other patients are traced. In this way, back-end data computations and the data architectures on which they rely steadily interfere with front-end interactions, shaping and at the same time being shaped by them. This innovative combination of computational and networking solutions sets *PatientsLikeMe*, and perhaps recent social media platforms more generally,[60] apart from other forms of collaborative networking supported by information and communication technologies (Benkler, 2007; Faraj *et al.*, 2011; Majchrzak *et al.*, 2013; Treem and Leonardi, 2012).

Some might find this conceptualization unsurprising. At a very basic level, all networking services of information and communication technologies depend on computational operations. Routers and switches coordinating the flux of networking data through algorithmic computation and e-mail clients receiving and sending e-mails are typical examples. Clearly, at this general level, the interpenetration of networking

[60] To those familiar with social media such as Facebook, our explanation should draw to mind various fundamental features such as the 'Like' action.

and computation is intrinsic to the current technologies of computing and communication.

However, our claim concerning the mutual implication of networking and computational capabilities is evidently much more specific. The links between patients and the patient activity in the network driven by those links are organized through a wide range of connections between patient data and patient profiles that the system is able to compute by relying on advanced data techniques. Database operations, we claim, lie at the heart of this *computed sociality*, as it were, which is realised by means of advanced representation and aggregation techniques that ceaselessly construct links between network members, here patients (Alaimo, 2014; Van Dijck, 2013). These observations of ours suggest that social media platforms are not vehicles of unconstrained socialization but complex technological arrangements that recast sociality in a network of social affinities that are shaped by computational operations. As we have shown above and in the empirical findings, patient interactions with one another and with staff are to a significant degree mediated by continuously updated links between network members previously unlinked and unaware of each other. It is this dynamic and constantly updated linking of patients to other patients via the intermediation of scores, counts or categories that shows the complex technological underpinnings of the network and makes it and similar ventures innovative and theoretically interesting (Kallinikos et al., 2013).

### *Institutional Implications: New Arrangements and Forms of Medical Work*

The mutual implication of networking and computational operations generates the need for a specific kind of expert work, performed in the process of symptom

mapping. The openness of the data collection to phenomena of various origins means that the system collects information on a much broader range of circumstances than traditional approaches would allow. This includes recognized medical entities such as symptoms, treatments, and conditions, mapped on a continuous and longitudinal basis. It also entails, though, as the symptom creation process demonstrates, data on everyday experiences and events that evade prescribed categories and, not infrequently, test the boundaries of what is, or may become, relevant and meaningful.

For patients, tracking everyday experiences can represent opportunities to communicate with other, similar patients, along dimensions that they find meaningful or worthy of pursuing – in addition to the possibility of personal health record bookkeeping. What is captured in the system of representations becomes a matter of convergence or divergence between experiences and life histories. By adding a symptom to her profile, the patient establishes the sameness between her experience and those of many other patients. The patient converges towards others via the intermediation of a standardized reference of experiences. Alternatively, by creating a completely new symptom, the patient marks the uniqueness, or difference, of her own patient experience from that of anybody else. Through the creation of a new category, the patient creates an experiential signpost through which other patients might start to connect.

For the organisation, data of this sort represent the potential for making clinical discoveries, and identifying and storing meaningful information on medical phenomena and events that could otherwise be difficult to detect. Due to the idiosyncratic, ephemeral, or mundane character of many patient observations, turning these data into

something meaningful depends on laborious, expert work. As the symptom disambiguation process described in the preceding pages shows (see Figures 3 and 4), such expert work includes interacting with the patients online, seeking to nail down the precise meaning and reality of patients' observations. In this process, expert medical staff link patient observations to medical categories and definitions whenever possible. When it is not, they establish new medical items, which, once integrated into the routines of the system, will have their relevance tested by future patient observations and associations.

A few things are worth pointing out in this context. The symptom disambiguation and detection process occurs online without physical contact with the patient. By the same token, the process is mediated by verbal means and other communication cues, at the expense of bodily examination and the focus on bio-chemical markers. These things occur in an environment marked by the absence or, at any rate, the minimal presence of the emblematic figure of clinical research, the doctor. In *PatientsLikeMe,* doctors figure as data collection architects and researchers. They influence data collection through activities such as research projects, participation in the system's long-term strategic planning, and leading frequent, internal, data collection process review meetings. Clinical professionals such as nurses and pharmacists conduct the expert work of data integration that we have depicted. Where technology alone can suffice to provide them support, patients are independent, namely in reporting, selecting, and recording their experiences in standard forms. In exceptional circumstances, the system requires the labor-intensive intervention of clinicians to collaborate and control the completion of the data entry process according to organisational standards. A new kind of division of labor is thus established whereby

the tasks underlying medical research are differently distributed across the range of clinical professionals. Also, the alternative architecture through which data are collected transforms the very shape and nature of this work. While it is hard to assess the stability and practical embedding of these changes, the pervasive nature of social media across the social and economic fabric suggests that they may well be part and parcel of wider institutional and organisational changes (Benkler, 2007; Faraj *et al.*, 2011; Majchrzak *et al.*, 2013; Treem and Leonardi, 2012).

While the process of symptom mapping is often laborious, requiring extensive forays on the part of staff into medical knowledge (e.g. classification systems and definitions), it is essentially aided by computational facilities and advanced database and representation techniques. Exploiting the editable, open, interactive, and distributed nature of digital data (Kallinikos *et al.*, 2010), these computational means and resources enable the expert to draw links between varying phenomena. In many respects, this expert work is data work as Zuboff (1988) depicted it some time ago (see also Kallinikos, 1995, 1999). Of course, as our study shows, the social and technological conditions through which data are generated and analyzed have shifted dramatically since the publication of her influential work. However, the nature and implications of the work processes Zuboff associated with work environments infused by a variety of disembodied data tokens, the challenge of what she called 'mastering the electronic text' (Zuboff, 1988; ch. 5), persist. In some respects, the changes we have outlined in this paper suggest that the work environments Zuboff perceptively described two and half decades ago have become even more pervasive today (Borgmann, 1999, 2010; Kallinikos, 2010).

The involvement of broad audiences, enabled by social media platforms and web technologies (Zittrain, 2008), is the driving force behind the changes we have sought to depict in this paper. Crucially, the changes we refer to extend beyond industrial or routine work settings, and concern expert work and the processes through which one of the most emblematic of expert pursuits, namely the construction of medical knowledge, is carried out. The punctuation of the patient everyday, the mapping of patient experiences, and the wide reach of phenomena *PatientsLikeMe* is able to access are all made possible through vicarious descriptions, and the medical entities they represent. In this process, social (patient) data become the raw materials transformed into medical facts through the series of operations we have documented in this paper. As shown in the example of symptom data collection, the clinical professional can manipulate links between entities through data actions such as coding, merging, splitting and so on. Specific, advanced data techniques underlie these operations that would otherwise be so demanding as to render their execution unfeasible. Technology and the data management techniques it embeds underpin the routinization of a range of fundamental expert operations through which patient data are transformed into medical facts. These are no meager changes.

## Conclusion and Suggestions for Further Research

In this paper, we have studied the processes through which a social media platform, *PatientsLikeM*e, draws on patient self-reporting to pursue medical research. Using social (patient) data for scientific purposes is, in many respects, an extraordinary accomplishment. The production of medical knowledge has commonly been based on collective processes in which professional skills in data generation, analysis, and

validation have figured prominently (Bowker and Star, 1999; Timmermans and Berg, 2003). In medical research in particular, these processes have taken place in a dense institutional context characterized by established organisational arrangements such as hospitals and health care units and the modes (routines, tasks, standard operating procedures) by which such formal schemes operate. The social media platform we have described in this paper sidesteps these fundamental conditions on which medical research has relied, and provides an alternative path to medical, and more generally expert, knowledge creation.

A network such as *PatientsLikeMe* embodies organisational developments that escape the dichotomies of industrial versus grassroots organisations, and formal versus open, life contexts. It has been pointed out that innovations facilitated by information and communication technology enable 'greater organisational and institutional reach' (Clarke et al., 2003: 162). Also, these innovations power heterogeneous initiatives of knowledge production on the part of groups such as patient advocacy organisations (Clarke et al., 2003; Marks, 1997). Thus, it was correctly foreseen that *'the heterogeneity of knowledge sources can be interpreted as disrupting the division of "expert" versus "lay" knowledge and enabling new social linkages'* (Clarke et al., 2003: 177). However, the case of PatientsLikeMe attests to the coming together of "expert" and "lay" actors through the interconnecting facilities of a new socio-technical system. What this seems to suggest is the advent of the lay actor not as a challenge or substitute to the expert in the production of knowledge, but as a stable collaborator – as an operator upon which expert organising depends.

At present, it is difficult to assess the stability, promise, and possible drawbacks of the web-based arrangements we have studied here. There is no doubt that the access to the patient everyday that social platforms such as *PatientsLikeMe* facilitate carries significant promise for making use of facets of patient reality and experience that have so far remained beyond the reach of medical practice and research. However, there may too be drawbacks associated with professional turf battles and social conflict (Abbott, 2001). It is also difficult to ignore the suspicion that something important may well get lost when medical expertise is cast in the role analyzed in this paper (Bowker, 2005; Dreyfus and Dreyfus, 1986; Zuboff, 1988). These important questions necessitate further research into these alternative modes of pursuing medical knowledge and their implications. In this paper, we have sought to carefully document the *terra incognita* of pursuing medical research via social media platforms and patient self-reporting. While the precise resources and solutions by which such a task will be pursued in the future may vary, the need for documenting patient experience through the means offered by social media platforms and web technologies will persist and possibly grow. The diffusion of these social technologies across the social and economic fabric suggests that they may well be part of wider cultural change in which the boundaries of institutional and organisational practices and arrangements are refigured (Benkler, 2007; Faraj *et al.*, 2011; Majchrzak *et al.*, 2013; Treem and Leonardi, 2012).

Extending previous research on social media platforms and drawing on our empirical evidence we have been able to further theorize on the nature of these social technologies. Social media platforms, we have claimed, are not solely places of congregation (socialization) but of aggregation as well. A variety of data is constantly brought into new configurations via aggregation techniques, producing new

possibilities for interaction that, in turn, feed back into the data generation process (Tempini, in press). Not much is currently known about this computational, as it were, rendition of sociality (Kallinikos, 2009) mediated by back-stage operations in social media platforms (Alaimo, 2014; Van Dijck, 2013). The social relevance and realism of social objects (e.g. averages, aggregates) constructed by statistical operations has been a pervasive theme in contemporary scholarship (Bowker and Star, 1999; Desrosières, 1999; Hacking, 1990, 1999; Porter, 1995). It would be interesting to draw on these path-breaking works of literature to reflect on the ontological nature and implications of a sociality that is considerably mediated by computational means and instrumented via social media platforms.

# Till Data Do Us Part: Sociality and the Proliferation of Medical Objects in Social Media-Based Discovery

**Niccolò Tempini, LSE**

**Abstract**

In this paper, I set out to understand approaches to medical research which rely on social media to supporting discovery and participatory design, with a focus on the key aspects of data representations and the construction of sociality. I report on the observational case study of *PatientsLikeMe*, a well-known platform in the health sector that has been pioneering the use of social media for producing scientific research. The organisation develops the technology to engage and govern a user base of patients who contribute data on several dimensions of their health. The main focus of these developments is the data structure of medical categories that allows the organisation to construct data aggregates for research and to connect with distributed patients; and allow patients to connect with each other. Recapitulating the main evolution of the platform in its first 5 years, I highlight how social media technology ambiguously supports the scientific enterprise of knowing and empowering patients and their experiences. The continuous evolution of categories and resulting data aggregates, which I characterize and define with the concept of *social denomination*, continues changing the organisation's understanding of what kinds of subjects the patients were and what they were experiencing. At the same time, social media platforms give new status to the patient experience, which the patients are now empowered to express in a proliferating wealth of categories. The central argument of the paper is that while the

emergence of alternative and more open organisational forms is a welcome development we need to ask the question of whether an alternative sociality that involves patient users more inclusively through social media is possible.

## Introduction

The recent rise of social media technologies has without doubt greatly fascinated both academia and the general public. Social media has accompanied innovative social and organisational experiments (Aaltonen and Kallinikos, 2013; Benkler, 2007; Howe, 2008; Shirky, 2008, 2010) in a variety of forms which we still need to understand fully. These platforms have allowed unprecedented numbers of users to interact with each other, share their experiences (Boyd and Ellison, 2008; Treem and Leonardi, 2012) and access information resources generated by the network which had otherwise been very cumbersome if not impossible to create (Faraj *et al.*, 2011). They have empowered users with new communication media, their underlying designs often being flexible enough to serve the needs of both the occasional consumer looking for entertainment and recreational interaction and the knowledge work professional looking for networking and specialist resources (Benkler, 2007). Social media have continued to support the emergence of new forms of community life and sociality (Bowker, 2013; Kallinikos & Tempini, in press). These followed along the lines traced by early experiments of mass online computerization (Feenberg, 2010; Feenberg *et al.*, 1996), only on a different scale, both in terms of mass involvement and in the number of domains of everyday life that are affected.

While users go about their everyday affairs and their work, social media infrastructures routinely collect behavioral traces and user generated content in the form of digital data (Aaltonen and Tempini, 2014; Morris, 2012). In different ways, most of the companies managing social media platforms put information production at the center of their business process (Aaltonen and Tempini, 2014; Tempini, in press). Digital data are the raw matter that needs to be worked on for marketing information products or services to clients. Organisational processes and routines are shaped around this goal, characterizing governance of innovative organisation forms (Aaltonen and Tempini, 2014; Gerlitz and Helmond, 2013; Tempini, in press; van Dijck, 2013). Great corporate empires have been established along variations of this canon, and certainly a topic of discussion has been the new extended reach by which systems of economic valuation and exchange have penetrated the social fabric to capture and trade computable data packages (Cheney-Lippold, 2011) – aggregate abstractions of people's everyday practices and activities (Kallinikos and Tempini, 2011). Social media platforms are environments that are constantly engineered in order to produce information (Cheney-Lippold, 2011; Tempini, in press; van Dijck, 2013) out of the experiences that users cannot easily value otherwise (Hayek, 1945) – except with their own self. Organisations relentlessly tweak and innovate the social media codebase in order to improve their grip on the data flows, narrating the social processes they want to govern or tap into. The sustainability of their business models depends on it.

Most generic social media platforms such as Facebook and Twitter typically exploit their information resource to sell advertisements based on such digital traces, translating the rationale of traditional mass media advertisement models to exploit the new capabilities of digital infrastructures in order to define and package an advertising

audience (Aaltonen and Tempini, 2014; van Dijck, 2013). Other platforms have been pursuing more disruptive visions. A host of renowned companies have been pioneering scientific and commercial research through the involvement of a massive user base and the analysis of the data generated through the social media platforms they maintain. One of the domains these efforts have been most concerned with has been the biomedical (Prainsack, 2014; Topol, 2012), a most effervescent one due to its strong social and economic incentives for scientific and technological solutions to health problems.

One such organisation is *PatientsLikeMe*. The organisation, founded in 2004, is well known for having pioneered innovative experiments with social media technology, carving itself a unique place in the health sector that allows it to relate to the non-profit world of patient activist organisations and health care communities (Epstein, 2008; Rabeharisoa *et al.*, 2013), as well as the industry complex of pharmaceutical companies and providers of various healthcare solutions. *PatientsLikeMe* offers research services including access to "raw" patient data and custom research data collection and reporting.[61] As the unique, proprietary resource fuelling its research projects, the organisation exploits the social media infrastructure to collect health data from the patient user base. Patients self-report their health status in a number of health

[61] The for-profit organisation, headquartered in Cambridge, Massachusetts, connects more than 250,000 patients through its social media platform. Most patients are suffering from chronic or life-changing diseases, as a long-term investment in a social media platform is more meaningful to those that need to live with a condition. Using patient self-report data, the staff have published some 37 outputs between peer-reviewed journal articles, reports, editorials and others. There are a broad range of topics, including distribution across patient population of symptoms (Turner et al. 2011; Wicks 2007) or psycho-social issues (Wicks and MacPhee 2009), patient education by medical practitioners (Wicks and Frost 2008) and virtual clinical trials (Wicks et al. 2011). More information at http://www.patientslikeme.com/about.

dimensions (conditions, symptoms, treatment, lab measures, hospitalizations, and others), constructing a longitudinally rich journal. Patients can report their experiences by either aggregating their data with other patients' data under already existing categories, or they can initiate the creation of new categories describing phenomena that have not already been reported in the platform. The architecture of data collection is data-based but also open to the patient voice in the form of user-generated definitions (and other kinds of content such as patient forums).

The platform in turn is able to connect a patient to other similar patients, filtering and ordering the patient user base according to the pieces of data that a patient has shared. Data structures are here *'gateway technologies'* (Ribes and Bowker, 2009:201) but turned over social interaction in order to connect individuals with others. The better and more detailed data a patient shares, the more the platform is able to compute, construct and display connections with similar known or unknown patients which are displayed across the website in its various pages (see Kallinikos and Tempini, 2014). At the same time, the better and more detailed data a patient shares, the more the organisation is able to produce information that fuels the research the organisation conducts and sells to clients (Aaltonen and Tempini, 2014; Tempini, in press). The generation and collection of informative data is therefore the main source provision activity of the firm. An organisation that, beyond merely supporting online interaction, aims to produce medical research through such architecture, entirely depends on a stable and rich supply of data; and faces specific challenges that contrast with other examples of distributed science or online communities (e.g. Bowker and Star, 1999; Leonelli, 2012; Millerand and Bowker, 2009; Ribes and Bowker, 2009; Ribes and Jackson, 2013).

Patients are involved through online communities in the production of scientific knowledge, and consequent business value, through the sharing of their experience and of the patient voice. The organisation has aimed to disrupt the incumbent industry complex (Clarke *et al.*, 2003) through epistemic practices that make the patient voice heard and accordingly shape research agendas and funding. The promise is to improve medical science and healthcare towards a renewed attention on the patient experience and testimony (Rabeharisoa *et al.*, 2013). This innovative socio-technical arrangement, dependent on the development of a suitable social media platform intermediating the data-sharing activities and making accessible to patients a wealth of health information that aims to engage the patients and compensate for their data sharing efforts, depends on the technology to realize some important social transformation that appears to respond to these hopes (Callon, 2009; Feenberg, 2010). In this respect, two major and interconnected fields of urgent inquiry appear immediately to the fore, which will be the center of this investigation.

***Social Data, Sociality and Representation. Or, 'Who are they?'***

The first topic of investigation of such a socio-technical configuration concerns the structure of social relations between different actors, as the social media platform affords them to unfold while itself evolves through cycles of design, development and adoption. This has been referred to as the loop between the first and second technology instrumentalizations in Feenberg's instrumentalization theory: it is the problem of how the experience of end users can feedback into the construction of technology and the understanding by the developers of what are the uses, purposes and value that the

technology embodies. Instrumentalization theory describes the issue of the co-construction of the human and the technological within a two-level conceptualization. The first instrumentalization concerns the meaning and functions of a technology as conceived by its developers and managers from the specific perspective shaped by their life context: *'primary instrumentalization is the process of de-worlding inherent in technical action'* (Feenberg, 2010:150; cfr. Heidegger, 1977). In the concept of second instrumentalization, Feenberg instead imports the stance of science and technology studies (STS). The second instrumentalization concerns the reception and adaptation of technology in the life context of its intended and accidental user base (Feenberg, 2010:150-151). New meanings and functions are discovered and form the social identity of a technology (Ciborra, 2003; Faulkner and Runde, 2009; Feenberg *et al.*, 1996; Orlikowski, 1996; Prainsack, 2014), feeding public discussion, informing regulators, creating markets (i.e. AppStore) and shaping future iterations of technological development. Social media platforms are typically flexible architectures that can - subject to minor variation - adapt to the needs of great varieties of users. The constant and increasingly easy updatability of the codebase has allowed social media organisations to let users get to grips with technology, make meaning, discover affordances and invent workarounds that could be later integrated in the formal offering of the social media platform (Faraj *et al.*, 2011, van Dijck 2013). Design processes have been opened and have integrated, with increasingly frequent cycles, the feedback from the real world experience users make. This has supported arguments for a more inclusive understanding of technology development (Feenberg, 2010).

However, as I will demonstrate, in these burgeoning social spaces where social interaction is intermediated by the generation and transaction of data, it can be a

daunting task to understand how users are related and aggregated to social groups. This is an important issue. The most famous examples of social media tend to be technologies developed for generic audiences (e.g. van Dijck, 2013), but in networks with a specialist focus, different kinds of social groups can converge with very different needs and experiences (Shirky, 2008; Weinberger, 2007). In the example of the research network *PatientsLikeMe*, patients share and feedback their data and experience as members of condition communities or other kinds of patient groups sharing a health-related focus. These seemingly fluid configurations and social groups could overlap to some degree with "offline" interest groups of similar roots (Epstein, 1996; Feenberg *et al.*, 1996). In such a peculiar convergence, a paramount question concerns the provenance of the (patient) user experience that feeds back (second instrumentalization) to social media developers and managers, and the feedback process by which it gets to them.

As social media are technological environments that engineer sociality by relying on categories, labels and numbers to intermediate social interaction and representation (Kallinikos and Tempini, 2014; Tempini, in press; van Dijck, 2013), to understand how real world user experience is feeding back to developers and designers requires tracking the evolution of such cognitive devices through the management by the organisations controlling the iterations of platform development. Data structures are *'gateway technologies'* - Ribes and Bowker use the term to describe software ontologies in distributed science (2009:201) - allowing one context to be connected to another. In social media, data structures are also gateway technologies allowing developers and users to access other users.

Understanding the evolution of the data structures powering a social media technology allows to the opening of a perspective on the identity and foundations of online user bases, groups and communities. In such a purview, and given the constant updatability of digital platforms and the fact that online social connections often do not enjoy a parallel "offline" existence, the paramount question to ask is *how is a social media infrastructure shaping the foundations of online groups and communities, which it purports to serve?* And*, how are such groups and communities in turn shaping the evolution of the infrastructure through their experience?* In fact, if we are not able to identify how social media user groups and communities are formed and defined, whose experience is being fed back to contribute to the development of the platform and the fate of medical research projects? This is a matter of utmost importance if we want to understand the contribution of social media to social and political life. How the social media organisation comes to understand who the patients are is crucial in discussing the implications of its governance.

If we want to assess the broader relevance of social media platforms we cannot be content with generic arguments or bucolic metaphors like "the hive", implying absence of a center and a pure essence of unrestrained sociality (Kelly, 1996). For these concepts to be valid descriptions of social media phenomena they must be the result and not the assumption of attempts to capture the elusive transience of digital sociality. How are certain (patient) user experiences fed back and selected for improvement, and who are the successful user groups? We must make an attempt to explain. Are social media just empowering pre-existent offline communities? It might be possible. Patients of chronic and life-changing diseases, compared to football supporters, shoppers or digest readers, are easily defined and stable. However, this seems unlikely, given the

previous considerations of the development dynamics of social media platforms which stress their restless development and dependence on complex data models. We need to discover whether social media communities are founded under different premises than their offline analogue, and if such discovery is possible (Kallinikos, 1993). This is at once a problem of identity and representation.

***Social Data, Human Experience and Discovery. Or, 'What do they mean?'***

The kind of research approach *PatientsLikeMe* engenders, systematically involving a mass of non-professional patients for data generation and collection, has been praised both for the promise of great clinical discoveries and for the potential of disruption of the pharmaceutical and health services complex (Shirky, 2010; Topol, 2012). Epistemic practices seem to be ever more open to the participation of the wider "lay" public (Wynne, 1996) and research directions and agendas can potentially be shaped by patients (Rabeharisoa *et al.*, 2013), once only passive recipients of care and objects of research – here stable collaborators *'upon which expert organising depends'* (Kallinikos and Tempini, 2014). Clearly then, the issue to be researched in a social media research network is also about the epistemological status of the (user) patient experience in respect to the production of scientific medical knowledge. Patients share their life testimony through whatever language and system of signification they have available. How can their specific accounts matter? Here the concern is not as much with the fragmentation of data collection activities and their distribution across a non-professional population (Kallinikos and Tempini, 2014). Instead, I want to reach a deeper understanding of how patient data can rise to the status of medical evidence. Whatever the experiential meaning one wants to signify through the input of a data

point in *PatientsLikeMe*, this refers and connects to a life context that will remain largely out of reach (Kallinikos, 1993; Mingers and Willcocks, 2014). When the organisation's researchers try to make sense of patient data, the context that shapes the meaning of the same data point also depends on the other data that can be pieced together to try and understand. However, the question with social media networks, where complex infrastructures are developed and expanded on a rolling basis, is of the many different meanings that the data can take on as the technology evolves and changes. This is an important issue. Understanding the process through which the organisation gets to understand what the patient experience is like is a paramount step to understanding how social media networks might realize their potential for inclusive participation.

This second issue concerns the question *how does (user) patient experience acquire meaning in the self-reported data that are the central resource of this approach to medical research? How is social media technology making it possible for self-reported data to be the foundation of building scientific claims?* To understand under what conditions patient data have informational value for research requires uncovering how social media and the categories and classifications that power them create a context that allows for the patient experience to be made sense of, abstracted from its context of origination (Jacob, 2004; Kallinikos, 1999; Zuboff, 1988), aggregated to similar experiences as testified by others, and ultimately made to matter in the scientific domain and public life (Rose, 1991). This is also a question of identity and representation. It regards how the patient voice can be made to count in the domain of medical science (Prainsack, 2014; Rabeharisoa *et al.*, 2013; Wynne, 1996).

These are important dimensions that indeed need to be examined to provide a complete picture of the social relevance of social media. How is the patient experience as testified by the (patient) user placed in relation to existing medical knowledge? How is the social media infrastructure supporting the patient voice's emerging and disrupting the slow and encumbered medical research (Rajan and Leonelli, 2013)? Which kinds of experiences matter for scientific claims that might influence the perception of health issues and the ways these issues can be addressed?

These are questions that point at major issues, such as the relationship between human lifeworld experience and scientific knowledge. The modern discourse surrounding scientific activity has been that of an autarchic quest to discover, describe and govern nature's mechanisms, but ultimately, meaning and utility are rooted in the human world (Feenberg, 2010). To find a new way to direct medical research to meaning and priorities that matter from a human standpoint is a pressing necessity that PatientsLikeMe seems to exemplify in its singular position within the health research sector. The *PatientsLikeMe* project involves assimilating the lived experience of patients and putting it into relation with existing research bodies and frameworks. The potential outcome could be to shape medical research towards agendas that are closer to the patients' hearts, but also empowering the pharmaceutical complex with powerful, granular information services that offer a way to understand patient concerns. With this information, novel narratives can be constructed that aim at shaping a patient's self-understanding and experience, providing new frames for reflection and action in association with certain therapies and solutions (Rose, 2006). The aggregate outcome is difficult to predict.

On the one hand, *PatientsLikeMe,* with its distributed and data-based data collection approach, representative of approaches centered on accumulation, aggregation and permutation of data sources, seems to repeat a modern approach to science – looking for the 'mechanics' of nature. In a social media platform, the reliance on data structures for coordinating data collection and aggregation seems to suggest that discovery cannot but be limited to the iron cage of recursively analyzing those phenomena that more easily lend themselves to abstraction and calculation (see Bowker, 2013; Feenberg, 2010:198; Rosenberg, 2013). One can wonder how the social media platform can give voice to aspects of the patient experience that cannot be easily de-contextualized in data, if the power to count that it affords depends on the counting of things. We do not want to buy into overly simplifying views of the effects of "datafication" of the social world and associated technologies (e.g. Mayer-Schönberger and Cukier, 2013). We would rather keep in mind that *'computers may have the data, but not everything in the world is given'* (Bowker, 2013:171).

On the other hand, the data collection architecture upon which *PatientsLikeMe* relies, essentially open to the contribution of user-generated definitions and content, has allowed patients to both initiate and participate in scientific discoveries and, as I will show, to progressively highlight to developers how the platform could be improved. In this respect, the project clearly seems to realize some of the hopes for participatory medical research, exposing science to the multiple perspectives from patients of all walks of life (Kallinikos and Tempini, 2014; Prainsack, 2014). The research network is centred on an essentially open data collection architecture, allowing the patient experiences to be represented in numerous nuances.

It can be expected that the answers to these two kinds of questions are unlikely to be univocal. Various technological components and social actors might be contributing to the evolution of the platform in conflicting ways. Once again, it becomes necessary to unpack the workings of the platform, and the fundamental intermediation role played by data, in order to understand what social media technologies represent in this broad horizon. Until today, social media have been mainly resource-intensive infrastructures that require investment by a sponsoring organisation – be it financial, human resource-related, or both. The organisation needs to set out and execute a strategy that will afford the sustainability of the infrastructure. To trace the development of these socio-technical systems, pioneering new participatory architectures in critical domains such as medicine and science is of paramount importance. By examining the conditions of scientific participation in an open, distributed and data-based technological environment, we can confront the above questions with detailed empirical insights of the most recent and most sought-after social developments.

Answering these research questions is the concern of this paper. The exploratory nature, and dramatic horizon, of the questions I have formulated invites further discussion and comparison with other empirical cases. The paper is structured as follows. The next section presents the research design and methodology of the *PatientsLikeMe* case. In the following empirical section, I recapitulate the history of the platform in its first five years of development, highlighting the data management struggles that have shaped the system – which must cater at the same time to the needs of many and diverse patient contributors, and of the organisations' researchers. In the discussion section, I develop an argument that tries to demonstrate how the continuous

and incremental development of the data structures generated path-dependency effects that continuously changed the representation of individuals and groups, and generated a wealth of new medical objects. The argument concludes by implying that organising information production (and science-making) through social media faces specific challenges that are connected to the full technological intermediation and loose participation characteristic of these environments.

## Methodology

I set out to conduct an in-depth, intensive observational case study (Sayer, 2000; Yin, 2009), to understand how social media data, and the classifications and categories that give them form, shape organisation and are used to govern distributed user bases, select user behavior, and understand users and their life contexts. Embedded case studies are a powerful research tool when the researcher faces a lack of strong operational guidance from existing theoretical frameworks, or intends to explore and understand new settings and phenomena (Sayer, 2000). As I have argued in the previous section of this paper, social media technology is a recent development and still necessitates an investigation that opens up the technology and looks at the role of data, categories and classifications in the light of questions of democracy and knowledge politics.

I visited the *PatientsLikeMe* headquarters over a period of 26 weeks between September 2011 and April 2012 for the purpose of conducting an in-depth, intensive observational case study. Without any monetary exchange between the organisation and myself occurring, I participated in regular work activities, including several projects

where I acted mainly as a member of the R&D and HDI (Health Data Integrity, more later) teams. The nature of my involvement in the organisation was twofold, both of researcher conducting a case study and of researcher joining forces with the organisation in its system design and development efforts. I was granted access to information resources that a regular employee would have.

During the period of observation, I had the opportunity to collaborate with most of the employees in the company (30-40 members). I also interviewed most of them, sometimes twice, holding 30 interviews in total.[62] The interviews were held both in the early phase of the study, to bootstrap my introduction to the field, and in the last phase, to fully open up the themes, and cross-check and validate the interpretations, that I had been building during my time on site. To complete triangulation of different kinds of data (Yin, 2009), I also participated in meetings (numbering 128) – among others company meetings, project task-force meetings and release-demo meetings; participated in and analyzed e-mail communications; collected numerous documents, slide-shows and screen snapshots of both user-facing and admin-facing systems. I logged observations in the form of electronic notes at any time of day, using note-taking software and a laptop computer.[63] Many of these notes were at once useful for my case study data collection, and for the work I was involved in with the organisation. I continued constructing tentative interpretations of the events that I was witnessing during the data collection period, by writing reflections in my time off-site and in order

[62] The average duration of the interviews was 60 minutes. Interviews were transcribed and triangulated together with other written and documentary evidence.
[63] Snapshots and written notes amount to 665 analytical episodes. Obviously, they are of different importance and size of content.

to follow up in the planning of further data collection efforts (Aaltonen and Tempini, 2014; Sayer, 2000).

In order to understand the role of data, categories and classifications in shaping organisation, it was necessary to search for and uncover the processes in which they were expressing their power the most. Soon after the start of the fieldwork, I had identified the research and data integrity activities as the front of the organisation's understanding of how medical representations and frameworks embedded in the system shaped research and patient behaviour. The teams involved in these activities were also originating much of the initial concepts and requirements for future iterations of software development and evolution of the frameworks. For this reason I immediately focused on describing these processes and mechanisms (Avgerou, 2013) and the resulting techniques and means by which organisational actors were trying to achieve their goals. From this starting point I expanded my gaze towards other processes that were feeding back on the development of the platform and of the category and classification frameworks, such as community management and business development. Broadening the scope of the observation was necessary in order to capture the consequences of data forms and categories in organising business and research through social media, and to accordingly construct an empirical narrative that accounts for these phenomena holistically.

## Empirical Findings: The Architecture of Experience

*PatientsLikeMe* is a platform connecting more than 250,000 patients, suffering from more than 1400 conditions,[64] through its social media infrastructure. Patients self-report their health status, logging data over time about a number of health dimensions. They maintain their personal profiles, designed to provide a snapshot of the patients' present and past health developments through infographics, scores and text. To this end, the platform offers a number of health tracking tools capturing the patients' health status along a number of essential dimensions and through electronic forms and questionnaires. These include conditions (primary and co-morbidities), symptoms (severity, associations with treatments or conditions), treatments (dosage, form, frequency, side-effects), lab measures (for instance, blood cell counts or forced vital capacity), hospitalizations (reasons, dates), disease-specific PRO (patient-reported outcome) questionnaires, and others. To capture these dimensions it is essential to be able to construct a record that says something about the health and life experience of a patient and that can sustain the research in health economics and drug evaluation that the company has been developing for its clients.

The system automatically computes and renders web-pages that display scores and charts about one's profile, or the state of specific medical phenomena across the platform as a whole. In the first instance, pages such as *'My symptoms'* put the log of one's symptoms in context, by displaying a longitudinal view of the tracked symptoms and related severity. In the second instance, pages such as the *'symptom report'* page or the *'treatment report'* page give an overview of the information that has been shared

[64] Accessed September 30th, 2014. http://www.patientslikeme.com

about that particular medical entity on the platform, including the distribution of a symptom's severity as experienced by the platform members, demographic information about the patients experiencing that symptom, and the treatments that those patients take to fight it.

In addition to the structured health-tracking tools, the social media platform is replete with staple social media features such as private messaging, broadcasting, commenting, and spaces for self-representation (such as the profile picture, or the *'About me'* textbox). Patients also participate in the platform to connect with other patients, for socialization, support, for sourcing information about alternative treatment regimes or coping strategies, to know more about others' dosage of a drug, or health risk thresholds (Wicks *et al.*, 2010).[65] To this purpose, the platform includes, in addition to the *report* pages, a number of forum rooms and threads, organized hierarchically by condition type (plus a few miscellaneous forum threads for topics such as politics or platform announcements). Therefore, forum life is organized around, for instance, forum rooms for cardiac diseases, neurological diseases, etc.

Importantly, the technological infrastructure constructs, shapes and fosters patient sociality through particular technological solutions (see Kallinikos and Tempini, 2014). It produces and spreads a wealth of links to other patients through the platform, which are computed and drawn on a continuous basis based on pieces of data that allow the filtering of the user base. For instance, by browsing the report page for a symptom, a

[65] Perceived benefits of PatientsLikeMe by respondents of a survey included 'learning about a symptom they had experienced', 'understanding the side effects of treatments', 'find another patient who had helped them understand what it was like to take a specific treatment' (see Wicks et al. 2010)

number of links are displayed to a patient that can connect to other patients who are suffering from the same symptom, or are taking a particular treatment for it, or are talking about that specific symptom in the forums. Based on computational operations and data-driven linking, occasions for preferential interaction are drawn and served to the patient. The interaction possibilities that have been engineered into the *PatientsLikeMe* environment realize a loop whereby the more about their health status patients share with the system, the more the system is able to produce (through continuous computational operations) useful and up-to-date links to other patients or pieces of information that have been shared in the system from elsewhere.

***Conditions as Horizons: Tight Coupling of User and Patient Experience***

The platform was founded as a for-profit venture in 2005 in Cambridge, MA by three MIT alumni as a site for allowing Amyotrophic Lateral Sclerosis (ALS) patients to connect and obtain social and emotional support by sharing information about their own situation - including treatment regimes and coping strategies. The three founders were trying to do what was possible to save brother and longtime friend Stephen Heywood, an ALS patient. The website was the family's second major venture after having founded one of the most prominent foundations for ALS research, ALS TDI.[66] The platform supported only one condition, and it was designed based on the extensive knowledge the founders of the website had of ALS patients' life experience and needs. ALS is an extreme case of a patient experience. There is no cure for the progression of

[66] There has been much interest in the Heywood family's desperate fight with ALS in the media with multiple articles, a best-seller book (Weiner 2004) and the feature film 'So Much, So Fast' (Ascher and Jordan 2006) dedicated to the period before the foundation of *PatientsLikeMe*. More information also at www.patientslikeme.com/about and http://www.als.net/About-ALS-TDI/

the disease and patients remain lucid throughout the convalescence while their nerves progressively lose control of voluntary motion, in a short time span (2-5 years life expectancy after diagnosis). Also, ALS belongs to the category of very rare 'orphan diseases,' for which research is very difficult. For these reasons, ALS patients are known for their activism and interest to try and experiment with anything that might help to improve their health. The experience of this condition dramatically dominates a patient's life world and eventual co-morbidities recede into the background. At the time of the writing, some 6856 patients are on the *PatientsLikeMe* platform for ALS, an impressive proportion given that a very high majority of members are from the US and the total US ALS population is estimated at around 30,000.[67]

The forum was focused on discussing all matters related to ALS. Patients could track the progression of the disease through an ALS-specific PRO questionnaire, and a fixed list of only forty symptoms deemed by the staff to be the ones characterizing experience of the disease. Similarly, the system asked the patients to track a fixed list of treatments that the great majority of ALS patients take – above all Riluzole and Baclofen. The focus of the website on a single condition and the fixed lists built an implicit, clear context to the user experience: the site was about ALS. The list of symptoms to track implicitly associated the symptoms to the condition and its immediate consequences.

[67] See http://www.patientslikeme.com/conditions/9-als-amyotrophic-lateral-sclerosis for PatientsLikeMe counts. For global counts, see http://www.alsa.org/news/media/quick-facts.html. Accessed on 10th October 2014.

The founders soon realised that the data collection model based on fixed lists of items was too restrictive and wasted opportunities for learning. They decided to continue sponsoring a list of 'primary symptoms' and 'commonly prescribed treatments' for patients to track but also to allow patients to input and track other symptoms and treatments, opening the data architecture to folksonomy.[68] Quickly, patients started to generate a wealth of new symptom and treatment definitions, many of which were not medical, while others described the patient experience with more specificity than the expert system UMLS would allow (Arnott-Smith and Wicks, 2008; Tempini, in press).

The website grew slowly, adding other disease communities starting with Multiple Sclerosis (MS) and Parkinson's Disease (PD). The staff researched the conditions to determine what set of tools would be better able to capture the patient experience. They developed a PRO tool for MS because of the lack of an existing one, and they adopted a previously validated one for PD. ALS, MS and PD, very different conditions from each other, still belong to the same family of diseases (neurological). Some similarities emerge if they are compared with the rest of globally existing medical conditions. First, they are life-changing conditions that tend to take center-stage in the life of the patients and focus their minds Second, they are diseases for which there are relatively stable and easy ways to measure the impact of the disease as a state of impairment. Adding disease-specific PRO tools and with little additional adaptation,

[68] Folksonomy is one among several kinds of data structures for managing user-generated content. Data categories are created by users (this is often called 'tagging') and stored in the system for further content aggregation. In respect to other data management structures like expert classification systems, folksonomies are flat hierarchies. This implies that between two categories, no formal relationship can be inferred by the way categories are organized (see also Jacob 2004; Smith 2008).

*PatientsLikeMe* successfully managed to construct a patient experience 'metaphor' for the management of the disease, by evolving on top of the existing architecture and clinical metaphors initially developed for ALS.

However, patient accounts were tightly coupled with one disease. The disease-specific communities were each a walled garden of their own. Patients accessed the websites from different URLs, and the registration of their respective accounts was inseparable from the association with the corresponding community condition. The account was registered in association with a particular condition number that could not be changed at any point in time. Each number would drive the execution of different portions of the codebase. It would present the patients with a different condition history questionnaire — the condition history is the system's metaphorical equivalent of the clinical interview, where patients can state the date of manifestation of first symptoms, eventual family history, date of diagnosis, and other things.

Based on the condition number, the system customized the (patient) user experience, reconstructing a context appropriate to the (user) patient experience. The number was 'driving' the automatic association to the patient profiles of custom, disease-specific PRO tools. Other features of the system instead relied on the execution of code that was shared with other condition communities, but were customized by loading different configurations. On this basis, the system linked the patient to different forum rooms, customized the list of recommended condition Primary Symptoms and Commonly Prescribed Treatments, the system recommended data input at different intervals and calculated and displayed links to other patients accordingly, by parsing

subsections of the entire user base. The condition number was the central data point coordinating the system behavior.

Over the first five years of its existence, *PatientsLikeMe* continued along these lines, developing about 25 disease communities (including, among the others, Epilepsy, Fibromyalgia, HIV, Transplants, Mood conditions). For each one of them, the staff had to go through the extensive and time-consuming research that was required to assess what set of self-reporting tools would make for a (patient) user experience able to successfully align to the context of (user) patient experience. However, it was increasingly felt that the website could serve a larger global patient population as diseases are in the thousands. Expansion of the platform was slow, as development was dependent on partnerships with clients who would fund the research necessary for developing a community. The organisation was receiving thousands of requests a year from patients asking for a community for their own disease. Most importantly, it had become clear that the development process was in the long run unsustainable because of two fundamental aspects of the worldly patient experience of a disease that the system and its underlying architecture were unable to capture fully. These were the epistemological aspects of living with a disease, and the patient community life aspects.

The hardwiring of patient profiles and conditions meant that patient profiles were not designed to host co-morbidities. Conditions such as ALS or PD make the existence and impact of co-morbidities less central. But the simplicity of the model slowly became clearer when the staff learned how, often, patients struggle between MS and Bipolar disease, or Epilepsy. The complexity of real world situations made this model even more unfeasible. The huge number of combinations - of a condition with all

the possible co-morbidities - that might be meaningful from the perspective of a patient living in a particular life context were neglected by the current system's architecture.[69]

Some of these patients managed two separate, siloed accounts. Others quickly found workarounds to overcome the limitations of the framework and be able to track the phenomena of their experience that mattered. The organisation learned about the limitations of the medical framework by looking through the data into what patients were reporting. Soon after the introduction of the open architecture for symptoms tracking, which allowed the tracking of custom symptoms, the patients started to exploit it creatively, to more meaningfully accommodate it to their experiential context. Patients "sneaked" co-morbidities into the system through the backdoor, recording them as custom symptoms, trying to make the most of the framework functionality by aggregating condition data according to the symptom tracking data model. The result was a data collection problem, and deadlock, with tracking for "severity" of a condition on a None, Mild, Moderate and Severe scale - a clearly inadequate model.

The presence of conditions in the symptoms database that were the result of an inventive workaround was also allowing patients to link treatments to exacerbated, collateral conditions, because the treatment data architecture allowed the linking of treatments to symptom categories in two ways. A patient could report both the side-effect and the reason for taking a treatment. For both aspects, patients could pull and link categories from the symptoms database. From a medical point of view, the working

[69] For instance, a straightforward example from my informants was that HIV patients can be very vulnerable from a great amount of infectious diseases when their blood cell count hits certain thresholds, hence health tracking should not stop at tracking markers and symptoms for HIV.

structure of the framework was incomplete, as treatments can also cause the exacerbation of conditions or syndromes – and the system only allowed the linking of treatments to symptom categories. The work-around practice (storing conditions in the symptoms database) was then useful for understanding the patients' health situation, as conditions-as-symptoms could be linked to treatments - a more comprehensive account of one's situation.

However, the workaround was also producing a data aggregation nightmare. It was not possible to make the distinction between a symptom and a condition consistently and coherently communicate it on a systematic basis on the patient-surfacing pages, such as treatment reports and symptom reports. The system treated conditions in the symptoms database as symptoms, and computed and displayed the related data in a confusing and potentially misleading way. In this situation, a distinction could be drawn only by medically literate individuals. An informant recounted how in the "arthritis symptom page" (arthritis being a condition-as-symptom), on top of the list of treatments taken by patients suffering from arthritis was Copaxone, a drug that MS patients take for the MS condition. The prevalence of MS patients among patients reporting Arthritis (due to the relatively strong representation of the MS patient population in the network) was causing the context equivoque.

There were several other path dependencies caused by having started the development of the system with severe chronic and life-changing condition communities. These path dependencies were now undermining the validity of the framework for broader and more systematic data sharing. One problem was the absence of a way to record the end of a disease course. The system did not

accommodate the possibility that patients could recover from their disease, because the platform had started with incurable diseases such as ALS or PD. In the experience of other kinds of health situations the end of a disease is an important event that should be captured. Some conditions, for instance, can stop as a consequence of a transplant of a new organ. Back in the early days of *PatientsLikeMe*, patients recovering from a disease could only stop tracking or delete the disease from their profile (thus removing the record of having had such disease in the first place).

The context of an online community and sociality was also affected by the repercussions of codebase interdependencies. Forums were 'siloed'. Patient accounts, coupled with only one condition, were associated to forum rooms dedicated to this one condition exclusively. Patients could not access the forum rooms dedicated to conditions other than their own. While this siloing improved the quality of many forum conversations and allowed sensitive issues to be discussed more freely, there are aspects of the patient experience that one might also desire to share and discuss with patients suffering from other conditions. For instance, it could be useful, especially for patients participating in less established communities, to discuss wheelchair options with patients suffering from other diseases who are also using wheelchairs.

At these early stages of the platform's evolution, the layers of (patient) user experience and platform architecture were tightly integrated. A medical condition was the implicit background of the site user experience and social interaction, its encompassing horizon, rather than only a contextually-embedded object of experience. Figure 8 summarizes the old architecture of the system, taking three conditions as an

example, two relatively similar between each other, ALS and PD (both neurological) and one, HIV, very different from the others.

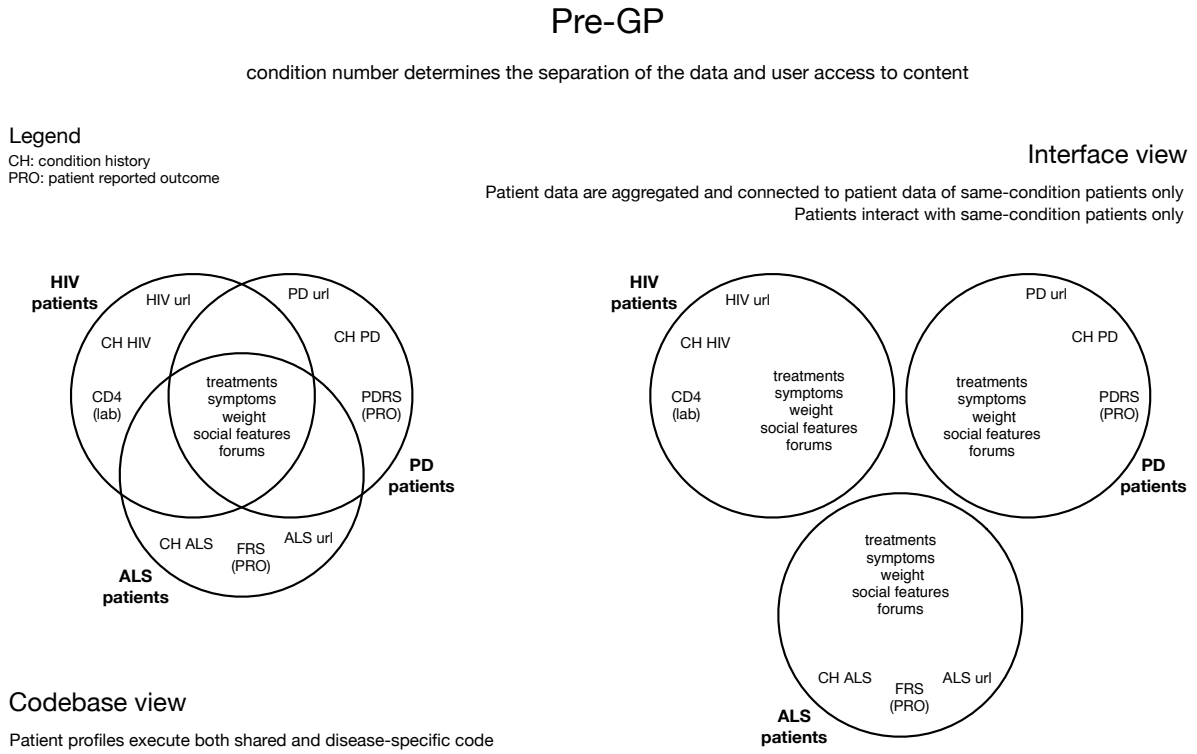



***Figure 8 – System representation Pre-GP***

### *The Generalized Product (GP) project*

The Generalized Product (GP) project, implemented in the spring of 2011, became a key turning point for the organisation in the understanding of the relation of co-construction between the platform and the communities it hosts. The organisation embarked on an infrastructural renovation of the system that had no antecedent in its own brief history, employing all of its development resources (beyond regular maintenance) for more than six months. The team started to unbundle the layers that glued together the components of the system. It was felt that the medical framework architecture had to broaden its scope and become easier to manage and configure, in

order to adapt to the diversity of the patient experience that the website was attracting on any given day. The system had to flexibly adapt to the changes in the underlying medical architecture. It was not feasible anymore to hardwire it to fixed configurations of patient profiles.

In order to make the medical framework more flexible and able to adapt to the broad variety of patient life experiences and contexts, patient accounts should no longer be tightly coupled with one condition. The system had to be able to capture the relationship between parallel conditions and the consequent ramifications on associated treatments, symptoms and other entities. Patient accounts and conditions had to become loosely coupled, able to establish and modify multiple relationships according to the evolutions of the patient experience. The system also needed to find a way to quickly scale the number of conditions represented in the system, and move away from the slow, in-depth and labor-intensive research process of condition modeling.

The GP was a new architecture of condition management, which allowed patients to have multiple conditions on their profile. To be able to quickly scale the number of conditions in the system in a way that actually served the needs of the patients, the new architecture was designed so as to allow the patients to add conditions to their profile, or request the creation of custom conditions. Patients could input a condition creation request when searching the database for a condition and failing to find it, in a way similar to how they had already been creating definitions for symptoms and treatments under the old architecture (see also Kallinikos and Tempini, 2014; Tempini, in press). Requests were reviewed by the Health Data Integrity (HDI)

team staff, who could then create a new condition in a few clicks. Based on a few configuration settings, the system would automatically deploy the functionality necessary to host the aggregation of patients of a new condition community. After less than a year, the system was hosting 1,400 conditions.

Under the new system architecture, conditions became an object that can be created and modeled at any time by the staff through admin tools. The once fixed condition number was translated over a new database condition_ID key, which would not act anymore as the overarching capsule of a patient community. Instead, the patients network became something much more dynamic and ineffable. Patients now fluidly belonged to groups of patients that shared a condition, in the context of a single, global platform population. All forum rooms were now open to patients, and the system interface automatically adapted shortcuts and links to forum rooms according to the conditions on a patient's profile, to make navigation easier and more meaningful. A patient suffering from both Epilepsy and Multiple Sclerosis would be encouraged to attend the forums for both conditions. Importantly, the various links to individual patients or patient groupings that the platform generates and distributes across the many webpages of the system in the form of clickable scores, icons and conversation snippets, could now be calculated by also taking into account the dynamic relationship a patient profile entertained with condition entities. Condition categories gained center stage as the main motor of the flexible architecture.

The new flexible architecture for the creation of conditions made necessary a number of further changes to other modules of the system. One such change was to make it possible to easily re-use the different health tracking tool modules that had

already been developed. A big part of the infrastructure development work of the GP project was involved in isolating a core of tools that would be available and shared by all conditions. Before GP, conditions used a shared codebase for some of the functionalities. With GP, the system features needed to be shareable on a needs basis. It was not sustainable to employ software engineering resources to reuse code. Through the new condition admin tool, conditions were created in the admin area of the site by the staff, by filling out an electronic form. The form generated and stored a configuration of system modules, and expert classification system codes, necessary to drive the behavior of the system and to aggregate the patient data for research.

Creating conditions on a scalable basis thus entailed standardization along a limited number of dimensions of the condition entities and the configurations of self-reporting tools and modules they are associated to. In the new system architecture, it was now not necessary to study every condition in depth before creating it. In preparation of the GP the staff had reviewed an authoritative list from the Karolinska Institute of more than 3000 conditions, and on its basis they had created a set of 6 condition types in which all conditions could fit: *Infections, Chronic Diseases, Pregnancy-related, Mental Health, Events and Injuries,* and *Life Changing Surgeries*. Each condition type was meant to drive different behavior from the system, such as the questions populating the condition history questionnaire and the intervals at which patients were asked to refresh their data.[70] Through the standardization of the questionnaires, the

[70] An informant explained how condition types afforded more appropriate condition history questionnaires: *There was Infections, for those things you could ask questions like 'When do you think you were infected? When did you first get symptoms? When were you diagnosed? Did you get a test?' Chronic Diseases, then you can basically just ask 'When did you first notice the symptoms? When did you get diagnosed? Are you taking treatment?'*

condition history would not entail a highly contextual, disease-specific set of questions, but would try to minimize the amount of unnecessary, off-topic ones.

By just filling in a name, a short name and synonyms, choosing a condition type and coding the condition against MedDRA, ICD-10 and SNOMED-CT, the staff created a new condition under which patients started to congregate and aggregate their experience data. The high majority of the conditions added in the first year since the GP launch were created with such minimal information assets. As the flexible architecture was designed to allow a highly diverse global user base to request the creation of new conditions, there was a need to monitor the patient requests in a systematic way. The speed of latitudinal expansion of the medical representations had been greatly increased, and the HDI team was coming to monitor a greater amount of data input. The new architecture included a dashboard for the staff, to allow the members to review the data input and take action. The dashboard displayed all the category creation requests that patients submitted whenever they wanted to track certain phenomena that they could not find already present in the database. The monitoring was necessary to preserve the purpose of data collection and their aggregation. Many patients indeed simply input alternative definitions of existing conditions. These patients' data should be aggregated to the other patients and their definition linked to the existing ones, so

---

*With Pregnancy-related Conditions: 'Have you had it multiple times?' Because you can be pregnant multiple times. And you would want to know 'When did you first think you were pregnant? Did this pregnancy lead to a live birth?' […] We realised there was a sixth type. When you have an organ transplant, you effectively acquire a disease called 'organ rejection' and you need to take immuno-suppressant drugs for the organ rejection. There are a few other conditions like that, like a bone marrow transplant and some forms of surgery. So we made them a sixth class called Life Changing Surgeries. We learnt that one on the fly as we were building GP.*

that their experiential evidence does not get lost and is made to count. In addition, patients often input as 'conditions' other medical entities, such as symptoms, or make simple errors - for instance, complex misspelling errors that automated tools miss to catch - that, if unguarded, might in the long run make the data needlessly fragmented. Other patients, instead, support disputed scientific statements – a topic that can be very sensitive. In all these instances, the help of the expert clinician is needed to settle the situation through knowledgeable judgment and a sensitive human touch in communicating with the patient.[71]

The HDI team also proceeded to a reintegration of all those conditions that had been inputted spuriously as symptoms, before GP, by patients that worked around a way to track co-morbidities. Over the years, patients had inputted more than 300 such conditions-as-symptoms entries. Many other patients had started to aggregate their experience data to the same condition-as-symptom categories created by other patients. Several of these categories gathered hundreds of patients' data. These spurious symptom categories were migrated into proper condition categories, but the symptom severity data (NMMS scale) was not translatable in the condition data framework. These historical data, of considerable longitudinal value, were "switched off" and practically lost, due to the lack of an infrastructure to compute and display them – a technology that would let the data "resurface" at the patient-facing interface. However, patients

[71] In the case of patients requesting disputed conditions, the team usually accepted the conditions about which evidence of a debate could be found in medical resources on the Internet, but made clear their status in the dedicated condition report page, stating it in the condition description. Through this routine almost only statements that are disputed outside of the system could become disputable inside of it. This is a delicate trade-off for a system that aspires, among other goals, to give space to minority accounts.

could now track their condition in a more correct medical context (as a condition category). Figure 9 highlights the change implemented through the GP project.

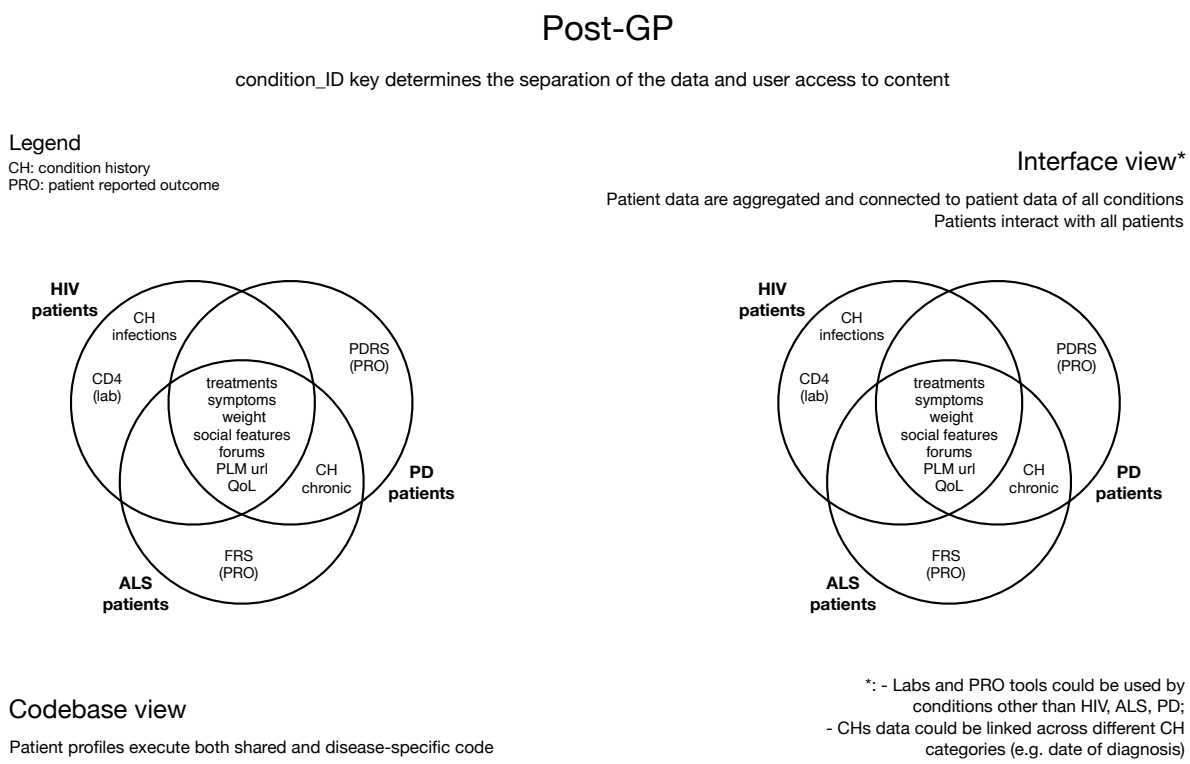



***Figure 9 – System representation Post-GP***

### *The Context of Community*

In the GP, while the medical framework was able, formally, to better reflect and adapt to the diversity of patient experience and its contexts, the user experience had become universal and generic, and the researchers felt somehow it was becoming increasingly difficult to know more about patients experience. Something was getting lost and data seemed less informative. The implicit context that the system had once been constructing, when it tightly coupled a simple medical framework and a siloed user experience, seemed now less coherent than before. One research executive explained:

*'When we built GP, I think what we really learned was that what we meant by a condition was something very different than the medical definition by a condition. What we meant by a condition was something that a group of people identified with. "I have this condition; I am this kind of patient." And there's a difference, and I think we didn't understand that distinction, and there's a big sociological element to a condition.'*

The upgrade allowed the system to quickly add numerous conditions and become a highly comprehensive health information web resource. Due to resource constraints, generally only the conditions of interest for a client-sponsored research project mobilized the resources needed for the team to research and configure what PROs, labs and other self-reporting tools could be assigned to patient profiles, and what disease-specific Primary Symptoms and Commonly Prescribed Treatments the system should ask patients to share experience data about. The user experience for patients for the majority of conditions was instead highly standardized. The system assigned five generic symptoms to all patient profiles, holding them as a basic denominator of the patients' experience of all conditions: *fatigue, pain, insomnia, anxious mood,* and *depressed mood.* All patients were offered the self-reporting tools for symptoms and treatments, together with the other general tools – the quality of life questionnaire QoL, the daily mood-health tracker InstantMe and the Weight tracker. Theoretically, patients' self-reported, open and distributed data collection could now be very efficient simply by the rigorous administration of this blend of self-reporting tools: conditions, symptoms,

treatments, and hospitalizations Correlational logic could suffice for the discovery of the *'hidden gems out there,'* as an informant executive explained.

The lack of customization of the conditions created with the new, streamlined process rendered an environment where patients were less active in the data collection and generally on the website. The general framework of the GP core was not enough in itself to sustain desired levels of engagement. The absence of new conditions for custom tools, very specific questions, and most importantly the community context that was implicitly constructed when websites were dedicated to one condition only, were making it harder to involve patients in what the website was meant to be about – community interaction and data sharing – and made it very challenging to attract and retain the critical mass of users that would make social interaction in the website a self-renovating resource.

After GP, many data points appeared now more clearly as disjoint from a firm context of reference. In an environment with multiple equally recognized conditions, researchers were unable to associate in a systematic way which symptoms were attributable to which conditions. Correlations were not sufficient in assessing these relationships, as it would have been too difficult to distinguish outliers from the 'rare gems out there' that the company was trying to be able to mine when patients track multiple symptoms, treatments and conditions together. The construction of a more assertive data collection context had often proven necessary to give patients some background to contrast their experience with. The organisation had experienced this since early on in its history. One informant recounted how one of the first publications to use *PatientsLikeMe* data (Wicks, 2007), uncovering a prevalence of excessive yawning

amongst ALS patients, required the explicit question to be asked to the patients. The system-wide questionnaire found 41% of respondents suffered from a moderate to a severe level, while before, the patients spontaneously reporting the symptom were less than a handful.

The relationships occurring between a condition and a treatment were also tracked sub-optimally, as patients could input a text string to report the purpose for taking a particular treatment, but they did not always cross-link this to condition or symptom categories, often inputting vague statements or non-medically relevant explanations. In addition, while the system was now able to record the difference between current and past conditions – this allowed the recording of patient clinical history over a considerable longitude – there was no clear way to systematically order the conditions by priority and tell the system what to prioritize. One informant elaborated:

> W*hy is a heart transplant more interesting than chicken pox? There may be a case where chicken pox is more interesting than a heart transplant. But what is it? Is it just that it is worse? What is the measure of worse? What's worse: stage four breast cancer or stage four pancreatic cancer? How do you measure that? Survival? Impact on life?*

Because of the GP update, the lack of context of the data was now emerging from the background, rather than being created by the software release. In the previous section, we have seen how patients, before GP, used to record conditions-as-symptoms, and consequently, the symptoms caused by these conditions, with no way to distinguish them from the symptoms they added that were caused by the 'primary' condition.

Before the GP, the context of the (user) patient experience was therefore uncertain. It was very hard to extrapolate what condition a symptom had to be associated to, and required expert judgment. However, the fast, effortless creation of many new conditions that were scalable solutions to the old data model straightjacket were in turn creating a user experience straightjacket - a "side effect" of the new medical framework that made available the same core of basic tools to address tracking of different disease experiences. Tactics for the management of the community aspects of the platform had become more generalized.

The way the system managed the forum rooms had to be rethought, to avoid the fragmentation of the patient community into a myriad of small discussion rooms where it would be ever more difficult to reach the critical mass necessary to maintain conversations and support redundancy of discussions (discussions about wheelchairs, for instance, could cross-over multiple communities). For the association of patients to the forum rooms (Skin, Hair and Nails; Endocrine, Metabolism and Nutrition; and so on), the architecture used MedDRA codes that were recorded by the staff in the condition configuration file. The use of MedDRA, an expert, symptom-oriented classification tree, for grouping conditions in larger categories, allowed the patients of hundreds of conditions to be distributed among about 20 forum rooms. These forum rooms were much more generic than before GP, cutting across multiple conditions at once.

A further generalization of the user experience involved the process of joining the platform for new patients. The welcome message and the other tutorial messages and activities now had to be generic so as to be applicable to multiple conditions. These

messages and interaction flows are designed to introduce and engage patients with the resources of the platform, but also to allow the system to learn about them so that it is possible to contextualize the user experience and make it more meaningful. While GP patients previously signed up through one of the secluded condition communities, thus implicitly confirming a most important aspect of their patient experience (being a patient of a condition), the generic sign up process (a consequence of the loose coupling of patient accounts and medical conditions) made the system unable to first guess what forum resources one patient might need, and provide contextual welcome messages and guides to resources that would get the patient started and connected with others. Providing a customized introduction to the website became a highly complex undertaking for which no automated solution was immediately available. The community team – the staff in charge of moderation and engagement in the forum rooms – was now unable to tell what kind of patients the new patients were. New patients were all registering through the same process, and it was not possible to know what conditions they suffered from (a good hunch about their needs and concerns) until they started adding conditions to their profiles. All patients were reverted to generalized messages, and the staff had to start re-thinking how to re-introduce customization at a latter point.

Despite implementing a more rigorous and mature medical framework, the GP had undermined the system's capacity to create vibrant and cohesive communities, fragmenting the user base into a few big groups and numerous small ones. Patients of new diseases, joining more generic forum rooms and a generalized set of self-reporting tools, obviously could not join a well-defined condition community. Patients of old, established disease communities instead were now also connected with other parts of

the platform, to which they would be associated by some 'secondary' piece of data. Patient search features, forum rooms, and other patients' data were now all open to all patients. Patients with severe and life-changing diseases were facilitated to interact with patients suffering from minor co-morbidities. The reception of these changes was different between the different pre-existing condition communities. However, some patients complained that they felt less oriented and able to navigate the website as they used to. In addition, opening all forums to all patients also increased the risk that patients from conditions easily associated to certain life styles and choices (think HIV) could be harassed by other members of extremely opposing views. The now piecemeal segmentation of the user base expanded avenues for sociality and made the framework for the generation of social links and interaction scalable and able to connect people across a broader range of medical phenomena, but along unpredictable and difficult to manage trajectories.

***Information... for whom?***

> *It wasn't until we added the GP, that I suddenly realised that she had breast cancer. Obviously, knowing that completely changes your interpretation of the rest of the contextual information.*

The interview quote is particularly "symptomatic" of the relentless and uncertain process of discovery and engagement of the patients and their life context. Understanding patients, the boundless combinations of their life paths and how to understand them through systems was the overarching concern in devising the next system's development iterations. The platform history in this narrative recapitulates the learning, by the staff, that patients generated and shared data that continuously

connected to a medical, social and personal context often remaining in large part beyond reach. Evolutions of the system had kept changing the understanding of who the patients were and what their experiences were. The data collected through the social media architectures that had been developed continually for years continued to further unveil the hidden connection of "facts", such as the statements that data appear to support if taken at face value with other, yet to be understood, meaning-determinants.

In face of all this contextual complexity that the system was increasingly making more evident and untamed, it became extremely difficult to devise systematic solutions for sharing information back to patients. How to know what kind of information patients need? An informant explained the new horizon that the GP had helped disclose:

> *Suddenly you go "Wow, this is not just about things we can even measure or things we can even know. It is about an incredibly wide variety of things." So, there are things that we can measure and there are things that will be un-measurable in the system.*

The platform's staff realised that in order to achieve ideal systematic information support to patients there would be so much to know about their life context, and an "exclusively medical" standpoint was not an option. The informant further reflected on the apparently simple problem of ranking drugs:

> *Because which HIV drug works best? Well that depends how rich you are? Do you live in Africa? Are you sick? Are you black? Are you drug resistant?*

Unable to capture all the facts and events that would possibly be relevant to understand patients' needs and concerns in the data, giving back contextual information

– contextual treatment efficacy, symptom distributions and severities – to patients proved a daunting challenge.

In order to develop a common point of reference for this new open environment of multiple diseases and infinite patient trajectories, a metric was developed that would be able to connect and associate patients along what they all shared: life. For this reason the Quality of Life (QoL) questionnaire was developed and added to the set of tools making up the core of GP. It comprised three dimensions (physical function, mental and social wellbeing), measured in four color bands (red, orange, yellow, green) and outputted an aggregate score that, in the hopes of the developers, would create a first, approximate common ground to compare the patient experience across diseases.

Understanding the experience of living with a condition through the intermediation of an open, social media space meant much more than dealing with the data collection of "sheer" medical data. It required holistically embracing the multiple dimensions and domains of human life. An informant explained: *'We should be able to think of conditions in multiple domains, they mean something medically, they mean something sociologically, they mean something from an importance standpoint, they mean something as a milestone in one's life.'* To conduct science for and with patients meant also rethinking what the platform should be about and concerned with. While the system had become more holistic in its representation of the global spectrum of conditions and medical entities, the GP had necessitated opening a different horizon for reflection, one considering the multiple dimensions of human experience. The informant elaborated: *'We confounded, we originally built communities that overlapped*

*with conditions, and thought they were about the condition. And then when we built a community of conditions, we did not understand what those communities meant.'*

Processing through the learning, the team continued with its development efforts to build solutions that would close the gaps they had progressively identified. The new underlying apprehension was that of the sheer complexity of the road ahead. The staff had more clearly discovered that it would become increasingly difficult to know more and discover precisely, through data, what kind of patient its platform members really were, what their experience was really like, what their concerns and needs really pointed at. While they had come a long way from their beginnings, their relationship with the patients remained bound to a certain degree of uncertainty.

## Discussion

Understanding the patients, their life contexts and experiences is the pivotal character of the processes of information production that are at the core of organising through social media for scientific and commercial research. At an immediate level of observation, it emerges from the empirical findings that the platform was governed with the overarching aim of improving the production of information (also Tempini in press). The events we have just recounted unfolded as the organisation incrementally developed and expanded the system to be able to produce more information about medical realities distributed in a world physically out of reach. The learning and unexpected issues that arose on this path were largely affected by the impossibility for the researchers to reach and access the very real world contexts from which the data are generated (Kallinikos, 1999; Zuboff, 1988). The organisation learned about patients

through data as the underlying data structures were changed and expanded. We can mark the centrality of information for this type of social media powered, distributed, open and data-based organisation by defining the work processes of expanding the functionality of the platform, renewing the medical framework and developing ever more flexible software architectures, as engaging in *information cultivation* (Tempini, in press).

The story of the evolution of the *PatientsLikeMe* architecture is rich with the emergence of ambiguities and complexities in the way data supported or questioned statements about patients and their experiences – the considerable success and hard work of the organisation in building a reputation within the scientific community and the wider public notwithstanding. The long and painstaking efforts spent on incrementally building a complex system that would help the patients to relate the data they generate to other pieces of information they reported, in order to reconstruct the context of their patient life, at several takes unveiled and apprehended underlying ambiguities. The evolution of the platform is then also one of the discovery of implicit and unexpressed contexts to which the data that patients had been inputting were inextricably connected (Bowker and Star, 1999). Ambiguities generated by the misalignment of the implicit context on the technological end of the system – the technology and its designers – and the implicit context on the human patient end of the system. Implicit context rested on the technological end of the system, as with the neglect of co-morbidities in the old one-condition architecture and the consequent stronger focus of forum rooms and communities. However, context was also implicit where patients operated, when they worked around this limited medical framework, starting to add conditions-as-symptoms, generating unconventional "condition

severity" data according to the symptom severity NMMS scale, and mixing the symptoms of co-morbidities in the same list with the symptoms of the "primary" condition. The workaround inevitably showed how the implicit relationship between "primary" condition and tracked symptoms could be easily confounded. The researchers would not be able to systematically say what condition each patient-added symptom had to be associated to.

Through rolling releases of the web-based application and most importantly by modifying the medical framework design and its underlying data categorization architecture, the organisation engaged in several attempts at redrawing the boundaries grouping patients and their experiences with others. Changes in the data architecture were reflected in the way data aggregates took form, as new versions of the system repeatedly "sliced and diced" the social data in new and multiple ways (Hacking, 1999). Major updates such as GP implied a fundamental re-thinking of what kind of subjects patients are in the system and how they relate to patients, conditions, and other medical entities. The updates radically modified how the system behaved and constructed digital links and spaces for patients to interact with other patients or track their own health and construct a self-narrative (Rose, 2007). The consequences of the GP implementation exacerbated feelings of disorientation. GP displaced established epistemic practices such as the co-morbidities workaround, and while the NMMS severity data that patients had collected for co-morbidities was not deleted, the new framework had made that data incommensurable and unusable. Being the cognitive pivot through which pass all the trajectories connecting social media objects and users to one another, the underlying data categories and aggregates are *social denominators* – here I take inspiration from the arithmetic concept of denominator which often

surfaced in fieldwork conversations. The progressive redrawing of the social denominator categories was aimed at the removal of data ambiguities. It reduced the misalignment between the implicit contexts of the technology and of the patient life, by operating modifications on the technology, the only reachable part of the system.

With *social denomination* (Tempini, in press) we can describe the observed technique of governance by which the organisation manages the platform and its productive relationship with its user base. The system evolved by relating the out-of-reach contexts of individual patients to data structures of progressively increasing complexity. Social denomination involved the exploration and production of new configurations of categories, which had to be significant for both the patients and the organisation. The denomination of the social is operated by modifying the structure of the categories and classifications - the cognitive grid that makes it possible to connect to the realities of distributed contexts (Kallinikos, 1993). Finding a common ground in a particular category supports, first, the imaginative act with which the patients can identify their own experience, communicate it and aggregate it to others' (Weinberger, 2007). Social denomination involves the creation and manipulation of new typonyms – or denominators, in the form of categories and classifications - to which the numerators (here the individual patient contexts) can at once oppose and identify. Second, the process of social denomination allows the features of complex health situations to be separated and understood at a progressively finer grain. It made it possible to produce more and more fluid datasets, where data could be translated and aggregated in a higher number of ways, for the construction of new meaning.

This has important implications. A technique for objectification and subjectification, social denomination flexibly displaces and redistributes, at each redrawing of the nominal boundaries between objects and organisational resources for the production of information. In this open and distributed form of data-based production (Aaltonen and Tempini, 2014) the data aggregate is the fundamental product of the platform at the center of the research efforts – a mold of social representations. If data gives aggregates substance, categories give them form. By repeatedly reworking the way people and their experiences are split and clustered together in different data sets, the organisation dynamically redistributed patient experiences in new data molds and linked them to changing collective identities. Evolutions of the system changed, on several occasions, the understanding of what kind of subject the patients were and what kind of experience their life was about.

In the introduction I identified two questions of urgent relevance in this new socio-technical arrangement for medical research. They concerned the issue of identity and representation in technology instrumentalization, and the tightly interconnected issue of the epistemic status and role of patient experience. I here recall these questions and formulate answers to them.

***Ephemeral representation***

In the outset of the paper I formulated the first question as one about the provenance and identity of the audiences that are engaged by the social media network, and in turn appropriate the technology and negotiate its identity. When distributed patient groups appropriate a technology that links them to a developing organisation,

we need to trace how it supports the process, and would like to be able to define how the identity of these groups is formed. In *PatientsLikeMe*, as it appears that the network affords considerable inclusion of social actors once marginalized from the research process, we need to answer to the question: how are patient groups formed through social media data aggregation representing their members and the interested others that lie outside of the network where the negotiation takes place? In a social media environment, we need to look first at how the technology supports the formation of said groups. As one might expect, the open and editable character of digital technology (Kallinikos *et al.*, 2013) makes the *PatientsLikeMe* infrastructure amenable to frequent restructuring of its architecture and inclusion of new tools and features. Further, as we might expect in an open and distributed architecture for data collection, the diversity of the user base elicited the creative re-appropriation of its function (Faulkner and Runde, 2009; Feenberg, 2010; Orlikowski, 1996). We have seen patients repeatedly testing the boundaries of a too restrictive structure of medical representations. A most clear example of this phenomenon (also Bowker and Star, 1999) was in the "conditions-as-symptoms" workaround. These phenomena are at the basis of the repeated engineering of the social media environment, which I have reported throughout the empirical narrative. The empirical evidence supports arguments that see social media as interesting and promising developments that open technology to politics different than those designed and to variable degree enforced throughout the development of modern technology (Beniger, 1986; Borgmann, 1999, 2010; Kallinikos, 2011). The contingent response of the organisation to the overflowing vivacity of online community life gathered to the *PatientsLikeMe* platform attests to a change in the way organisations can integrate the experience of (patient) users in the processes of technology design and development. This and similar platforms might be affording (patient) users more

influence than ever before in the shaping of complex information technology. In this respect, we must observe that the process of selective development of categories and data structures in social media networks, that I call social denomination, is also about assimilating feedback from patient users. Observing and learning from patient workarounds and the nature of the user-generated categories was essential information for determining further development of the platform. Social denomination is not sheer imposition. Instead, it entails a degree of learning from the information reported by the patients.

However, I want to also point at another phenomenon emerging from the empirical evidence that might be more difficult to integrate under mainstream arguments on social media and community. Social media are highly flexible technologies that invite reshaping and improving architectures and data structures on a rolling basis. Consequences of these activities of social denomination, as we have seen, are the repeated re-drawing of boundaries of things and social groups, molding them in new configurations with great flexibility and at several takes. However, data and the structures that give them form were not only the medium through which the organisation assessed the knowledge about patients and planned the next software development iterations.

The empirical evidence shows us that the organisation repeatedly discovered the ambiguity of notions of what kind of patients the users were, what conditions they were suffering from and what kind of life experience they were going through. Patients might be different people from whom they appeared to be by looking at the data. Consequent changes to the architecture managed to reshuffle patients and their data towards more

accurate representations and groupings, but at the same time the staff grew increasingly aware that exhaustive and comprehensive patient data collection would be practically impossible - for both modeling and patient compliance reasons. The role of data structures and aggregates in shaping social groups and representations were here playing at a deep and radical level, which is that the understanding of who the patients were continued evolving as the underlying infrastructure allowed the researchers to uncover further details and ambiguities. At several points, the experience reported by some of the patients suffering from arthritis came to be seen, in the pre-GP system, as that of MS patients, and in the post-GP system, as that of arthritis patients (subsequently, as patients of specific arthritis subtypes – as in Tempini in press). What I am trying to highlight, therefore, is how the continual evolution of social media platforms can be at odds with the notions of identity and representation that are of paramount importance in discussions concerned with sociality in new technological networks. To know who people are shapes the meaning of what they say – it makes context. At *PatientsLikeMe*, the staff had progressively discovered that it would become exponentially more difficult to discover ever more precisely, through data, what kind of patient its platform members really were, what their experience was really like, what their concerns and needs really pointed at.

***The Evidence of Experience***

The second question I formulated at the beginning of the paper concerned the epistemological status of the patient experience in respect to scientific medical knowledge production, on a platform that relies on social media architecture to make such experience relevant and translatable, through operations of data collection and

aggregation. To answer the question requires highlighting how social media and data structures create an information context of their own whereby one's experience, once abstracted from the context of origination, can testify elsewhere and conjointly with similar experiences by others without betraying its own essence.

Immediately, one can notice how the platform has broadly succeeded in including patients in the research process – I defined earlier as data entry operators 'upon which expert organizing depends' (160). The organisation has been successful in publishing research that depended, from an organisational point of view, on the voluntary collaboration of patients (see also Kallinikos and Tempini, 2014). Some of the successes even came about because of patient initiative, including the major one, an article disproving a drug's efficacy in slowing down the progression of ALS disease that was published in the prestigious *Nature Biotechnology* (Wicks *et al.*, 2011). In addition, it must be observed that the architecture of data collection, open to user-generated data categories in the form of patient requests for symptoms, conditions and other underrepresented phenomena, allows the system to include representations that do not figure in expert classification systems. When patients added conditions and syndromes of more debated status, the system was able to support multiple and competing statements on medical ontology. Most importantly, through data representations made of different layers, the platform was at once supporting and intermingling ontological statements about medical phenomena originated by, respectively, its staff (custom-made condition type categories – constructed for driving GP behavior), the patients (patient-generated synonyms and definitions), and the broader scientific community (clinician-validated condition configuration files, powered by expert systems coding).

The interdependence of the clinician, leading validation and design of the platform, and the patient, performing observation and reporting, defies attempts at characterizing this relationship as one of univocal hierarchy or parity. On one hand the relentless, subtle way in which patient experience fed back, percolating through the data and shaping the way researchers thought about the system's medical framework, testifies to a disruption of traditional models of medical evidence collection and associated clinician-patient hierarchical relationship. On the other hand, if we take a step back to a broader horizon we see that the platform was governed consistently with a vision that centered on the belief of making the patient world and experience accessible, transparent and discoverable through a specific technology-driven approach relying on complex data structures and data aggregation and correlation techniques (Kallinikos, 2009). There is a link here with the modern view of science and its associated formalization projects as value-free discovery of natural mechanisms determining the patient experience (Agre, 1992; Bowker, 2013; Feenberg, 2010; Heidegger, 1977). No wonder, as the organisation is dependent for the sustainability of its expensive business on the compatibility of its modus operandi with that of the wider industry complex it is inserted in (van Dijck, 2013). The reliance on computation for associating and comparing, through statistical regularities, different phenomena on the basis of their observable similarity is the spine supporting all of the most recent Internet-based solutions, from social media to big data, that have come to fascinate many of us and claim to subvert the way we look at the social world (Bowker, 2013; Kallinikos, 2012; Mayer-Schönberger and Cukier, 2013; van Dijck, 2013).

However, the prevalence of this specific perspective on the human patient experience, as the hidden background of the data collection, would perhaps be barely

noticeable if not for the impracticability of consistent access to the patient life context. It is not only that infrastructural changes could at once make unusable patient experience data that had suddenly become incompatible (remember the "condition severity" data on the NMMS scale that was lost with the upgrade to GP). It is also that the likelihood of a patient's lifeworld being fully captured through said approaches does not seem within reach. On the more practical side, the complexity and sheer number of relevant patient life events and objects that patients should report makes it nearly impossible to record them all, and especially so when the system is able to connect to the patient life context only partially. Indeed we have seen how patients were the creative force of the data collection only intermittently. *PatientsLikeMe*'s research on the prevalence of the excessive yawning symptom in the ALS patient population crucially depended on a researcher sending a questionnaire to patients, with the question made explicit (Wicks, 2007). On the more theoretical side we have long known that there are aspects of the lifeworld that simply are not amenable to being captured in the rigid data forms that computers store and process (Bowker, 2013; Feenberg, 2010) – '*Collectivities that are not being measured or modeled are preserved, if at all, only accidentally*' (Bowker 2013:170). We have seen how informants discovered this on their own. Other dimensions of the patient experience, such as the social and personal, and their worlds of community life and affects, were those on which the organisation had an admittedly weaker hold. The reflections on the essence of the patient communities, on the social dimensions to the evaluation of a drug's efficacy or on the difficulty of coding *"good interface on top of an increasingly generalized architecture"* that the GP project inspired, talk to us of the discovery of the cultural chasm often separating aspects of the patient experience that are strictly interconnected. The empirical narrative reminds us that for the several dimensions of the patient experience that the developers had tried to

understand and recreate in technology, the forum rooms and the community life they were supposed to open had been more difficult to formalize.

***Healthy boundaries?***

The inclusive, participatory process of creation of categories that allow the engagement, slicing and dicing of the social in multiple ways, and which underpins much of the social web, is a powerful and easily adopted one, as Weinberger (2007) has argued, because of how closely it aligns with how human cognition works and creates ordered representations of the world. The process involves the creation of hierarchies from the ground up, first through the association of a worldly phenomenon to a category, then by finding eventual relations to other categories from which more abstract levels of the hierarchy can be calculated by subtraction of features. This aspect of social denomination that includes categorization being opened to the audience can ignite a process of splitting and fragmenting of the social world that is potentially infinite. It allows the patients to separate their experience and identify it, on grounds of difference, from others' categories of experience (Bateson, 1972). This is an efficient way to keep the patients engaged, by allowing them to use a language that is familiar and close to their experience. In addition, with its combination with the operations of background mapping of the patient categories on expert classification trees, it can be a theoretically sound way to make differences, in the language of the data entry operator, while maintaining the ability to equate phenomena with each other at chosen levels of abstraction.

However, some other interesting consequences, from a medical point of view, are that all this fragmentation and further specification, while theoretically more precise, made it progressively more difficult to say with a degree of confidence who the patients are and who they are similar to. In this respect, the potentially infinite contextual-linguistic fragmentation, or breaking down, to which an individual patient case can be subject, seems to resemble, in its potential outcome, other developments in medicine that have supported a critique of the concept of natural normality (Rose, 2009). Rose indicates that the progress of genomics and brain science point to the progressive dissipation of those thresholds or characteristics that Canguilhem indicated as natural normality, a sort of baseline of the human – despite Canguilhem himself advancing along these lines in late work with a critique of the distinction between normality and pathology (Canguilhem, 2012). It is here compelling to suggest that there might be interesting similarities between those medical advances and the ideal outcome of the language-intensive and uncertain approach to medical discovery that I have been describing and which can potentially produce individuality as represented by unique configurations of data points in each patient profile.

At the same time, the open categorization process at the heart of the PatientsLikeMe approach, founded on a set of advanced computational techniques and data aggregation solutions, fosters the creation of a universe of micro-pathologies, symptoms and experiences of illness. In a way, this confirms that notions of the 'normal' and 'pathological' are linked to the data practices and techniques with which we have studied and made sense of medical phenomena (Löwy, 2011). In *PatientsLikeMe*, immediately after the creation of a new medical category, the system feeds the category back in the system and makes it available to all patients for data collection. The system

consequently aggregates patient data and computes new scores and counts. The patient-generated categories of these "micro-pathological" phenomena, which encapsulate their own "normal" in their signifier label, seem to give clearer and objective cognitive borders to new medical objects that are heavily grounded in culture and social practice (Rose, 2006). The immediate aggregation of associated patient data in counts and scores confers to these categories the objectivity that numbers have the power to give to socially disputed entities and claims (Porter, 1995). To these phenomena and through those data manipulation techniques, new trajectories of treatment can be immediately computed and associated, narratives can be elaborated and exchanged, and ultimately lifestyles and ethical corollaries about certain life experiences can emerge (Wynne, 1996). These data-based interactions contribute to shaping the lives of patients, reshaping their understanding of their own conditions, life expectations and treatment options, and informing their decisions.

The social media environment, by giving the patients the power to create new medical objects, and to give them shape through the data that they input and aggregate, acquires in this purview a very powerful character. It fosters the proliferation of objects, and the construction of their relations with other objects, which can shape the patient's life and also research, attracting in the data, as an executive informant once explained, "error, fraud, or really interesting stuff." Science is a social enterprise (Latour, 1987), and to open up science to new actors (Wynne, 1986) through social media, means to open it up to new forms of sociality, result of the (patient) user context as much as of the particular characteristics and operations that are at the heart of the technology that underpins the research network – giving life to new hybrids of scientific practice.

## Conclusion

To some, this latter development of the argument might seem too abstract. In this respect, it is important to maintain that the emergence of alternative and more open organisation forms is a welcome development, for the numerous reasons already mentioned. In the context of the health sector, it challenges a narrow vision of the clinician-patient relationship (Marks, 1997; Rabeharisoa *et al.*, 2013) and more generally a simplistic vision of the lay-expert divide (Prainsack, 2014; Wynne, 1996). However it is still important to keep in focus, first, the uncertain character of understandings of the human condition that necessitate reducing it to data categories and quantification and cannot escape technological intermediation. The article has demonstrated how daunting it is to define who the patients are and their experiences through a social media arrangement, and how the development of an increasingly complex system mostly exacerbated the apprehension of the complexity of a patients' life context. Second, the article has shown how social media technology, because of its underlying architecture, materializes very particular outcomes where radical, innovative and inclusive social agendas are put in operation through specific socio-technical arrangements. An organisation controls the development of the platform and needs to take specific decisions that are fundamentally driven by the need to learn about patients in a way that allows the production of salable information (van Dijck, 2013). Categories and data structures are shaped accordingly, in relation to the scientific and commercial uses they can support. At the same time, patients have considerable leeway in shaping the system and the forms of phenomena it tracks. They make proliferate in the system multiple and redundant accounts and definitions of their experience, constructing new objects to which specific forms of social life can emerge and are helped to spread.

This is clearly not an article that aims to promote a dystopian vision about the use of social media. I have defended throughout the article that the redistribution of power that social media makes possible is a welcome and promising development. However, the contribution of the article lies in dispelling simplistic thinking and the blackboxing of social media technology. The aim is to help a clearer vision emerge about what are the challenges of inclusion through social media, a promising but ambiguous partner for the scientific enterprise of knowing and empowering patients and their experience. This is also an important point for social media organisations, for which inclusion is a fundamental driver of success A number of further considerations are in place. It is important to notice here that the patient groups that could thus be drawn and shaped through changes in the way *PatientsLikeMe* technology works were not, apparently, reliant on the platform for engaging issues of direct or dramatic social conflict. However, my argument is still relevant in pointing out the limits of social media technology and asking whether different outcomes might be realised in contexts where social groups coalesce and mobilize around precise social issues. The imaginative contribution of fellow scholars would be helpful to connect this argument to other spaces of social challenge. In addition, it might be useful to compare the argument I present with other studies which analyzed earlier Internet technologies with similar agendas (e.g. Feenberg *et al.*, 1996). Most probably, earlier Internet technologies similarly shaped groups without warranty of a stable composition of members. However, one aspect that seems to separate social media technologies from earlier experiments is the combination of ever flexible technologies and data with the inclusion of said platforms in wider networks of production, which expose them to specific

economic incentives and organisational logics (Aaltonen and Tempini, 2014; van Dijck, 2013).

# Conclusion

## Overview of the papers

The papers have been presented in chronological order. At the same time, the order should reflect the incremental development of a tightly-knit set of arguments. In what follows, I briefly summarise the contribution of each one and how the papers are linked to each other.

### *Everything counts in large amounts: a critical realist case study on data-based production*

The first paper is a preparatory study that serves as pilot for the main study of *PatientsLikeMe*, concerning the case of an organisation that develops and manages a digital communications infrastructure affording distributed data collection from a broad and dispersed user base. The paper demonstrates how the sustainability of the organisation form depends on the steady and reliable production of information out of the large amounts of data that are routinely collected through the infrastructure. The data emerge as real structures of their own, coming to constitute the entire cognitive environment in which employees articulate their efforts, with the user base being largely out of reach. The reality of the data pool structure is at the centre of the argument. The article demonstrates that the data pool exhibits a set of emergent properties of its own that cannot be reduced to the individual data token components (comprehensive and unbounded, emergent; granular, resultant), and shapes the mechanisms of information actualisation that unfold at the organisational level.

The data pool is here raw matter of information production work, having a potential to produce information that requires specific mechanisms to realize. The potential of information is not fully contained in the data material. The organisation articulates efforts in order to improve the realisation of information from the data pool. These findings, the article claims, advance the understanding of many organisational settings that are centred on data processing. This perspective is integrated in the following articles on the social media case, where organisational actors are shown to be in continuous relation to the data and concerned with the assessment of their state. Highlighting the causal relations between data structures and organising, the paper also makes the case for the use of critical realism in the study of data-based organisations, and demonstrates the operationalisation of the framework and its techniques (social denomination; analytical writing), which have been then consistently applied throughout the other three papers.

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

The second paper examines the mechanisms of information production in *PatientsLikeMe*, to understand the conditions under which knowledge is produced through social media. The paper demonstrates the criticality for the organisation of the collection of data rich in information content, and the consequent importance of governing the user base towards information-productive interactions. In this respect the paper contributes to the understanding of organisational activities through the concepts of information cultivation, its mechanisms and the social denomination technique of governance of the patient audience.

While both the pilot and the main (social media) case involve organisations that must know their user base and keep users engaged in data-generating behaviours that are reflected exclusively through data, the important difference between the pilot case and the social media case emerges immediately. In *PatientsLikeMe* organisational actors are continuously assessing the data and trying to find ways to realise more information, but they manage and develop an infrastructure embedding much more complex data structures, and need to understand the network's users and their life contexts much more precisely and comprehensively than the MVNO employees do in the pilot case.

As a result, while the second paper is also concerned with information actualisation, the focus shifts from the organisational mechanisms towards the structural mechanisms of data-based information production. Different data models are shown to have different effects on the amount of information that the organisation is able to produce, not only because of how specifically they capture the world in a cognitive grid, but also because of differences in how flexibly they align with the contexts of adoption. The strategy of continuous evolution, or tinkering, with the data structures with the aim of improving the production of information is captured in the concept of information cultivation. The article also initiates an elaboration on the consequences of this fast-paced, restless process of redefinition of the objects of data collection with the concept of social denomination, a line of investigation that is continued through the following two articles.

***Patient Data as Medical Facts: Social Media Practices as a Foundation for Medical Knowledge Creation***

The third paper illustrates how the specific *PatientsLikeMe* data collection work architecture has been engineered to accomplish information production through systematic involvement of patient members. Firmly maintaining the perspective that sees data as key resource and the realisation of data's informative potential a critical concern of the organisation, the paper fundamentally shows that social media are not vehicles of unconstrained socialisation. Instead, it shows that social media systems crucially depend on data and data architectures for the realisation of both (patient) user value and (research) organisation value – contributing to the perspective that data and data structures are organisational devices for governing social media.

The article demonstrates how social media technology is leveraged to organise patient user base contributions in ways that break with the traditional configurations of data management practice involving medical research institutions and professions – which systematically marginalized the patients. The argument shows how specific technological solutions, based on web-based technologies and advanced data representation and computation techniques, make possible an entirely new architecture for the accomplishment of data collection work in medical research. It explains how technology allows the break down of data collection tasks into fragments that are manageable by patients alone without the constant supervision of the clinicians.

Also, and most importantly, the article shows how technology allows the construction of an environment where occasions for sociality are tightly interconnected with data generation and collection tasks. Through dynamic computation and

construction of links and other occasions for social interaction, the social media system automatically generates and renews incentives motivating desired user behaviour. The paper also shows that the result of successful patient involvement architecture is the new status that patient-generated forms for signifying patient experience gain in front of other, traditional, epistemic forms such as the rigid taxonomies of expert systems.

A major implication of this innovative architecture of data collection work concerns its role in relation to traditional institutions and professions and the established data collection and management practices that refer to them. In social media networks, a new kind of division of labour takes place, where tasks once prerogative of clinical professionals are now shifted to patients on a systematic basis. The data collection process entails limited professional involvement, and patients are largely unsupervised in the completion of data collection tasks. The possibility of executing data collection on a continuous basis, and from anywhere a patient has access to basic computing and networking facilities, also implies a breaking of the boundaries once separating the loci of clinical research from the context of everyday living.

***Till Data Do Us Part: Sociality and the Proliferation of Medical Objects in Social Media-Based Discovery***

The fourth paper examines broader social implications of the *PatientsLikeMe* social media-based arrangement for the production of information and medical evidence. It was a concluding claim of the third paper that some kinds of topics can be investigated only once research has provided a detailed account of the processes and operations that make the changes and developments of interest possible. The paper in this respect follows on the tracks that have been traced through the previous

manuscripts, albeit developing an original argument and adding further empirical insights. The empirical narrative summarises the first five years of the story of the platform and maintains a focus on data structures, the cognitive grids through which the system connects with the world phenomena. Through a history of the creation of the patient communities and the engineering of a system that needs to become increasingly able to adapt to all combinations of patient life contexts, it explains how the platform evolved in the way it did as the organisation tried to produce more information about more phenomena – information cultivation.

The article shows that social media-based arrangements for information and knowledge production reach deeper into the relationship between the human lifeworld experience and networks of value production. At the same time, the continuous development and tinkering of social media makes this technology an ambiguous partner for the scientific enterprise of knowing and legitimising patients and their experiences. Building on the concept of *social denomination* (first introduced in the second paper), the paper shows that the project of a more inclusive, alternative sociality needs to address issues connected to technological intermediation.

The empirical narrative demonstrates how the continuous shifting of categories and the resulting retracing of the boundaries and composition of social groups – operations of social denomination – translates into unstable social representation and grounds for sociality. In a system where, crucially, sociality and representation, experience and evidence are continuously standing on each other, *in vivo* operations of database development have deep implications. For one, they keep changing the points of reference that identify a certain individual along dimensions of sickness or health, of

patient of a certain kind versus another kind. The article shows that the social media practices of data management relentlessly redistribute patients and redefine who they are and were, trace new trajectories for online, computed sociality, and give new meaning to the broader context of each individual patient's data. Result of the open architecture of data collection is a proliferation of objects, relentlessly constructed and corroborated by the system through data aggregation, to which patients can contrast and link their own self-understanding, life trajectories, treatment options and decisions. As such, computed sociality is here shown to be an inherently open and undefined project, the dimensions and horizons of which are difficult to stabilise and share with a larger public, but that has consequences that must be unveiled.

## Recapitulation and Final Remarks

The thesis, by a comparison and contrast between a mobile virtual network operator and a social media network for patient socialisation and medical research, advances the study of innovative organisational forms revolving around the production of information through open, distributed and data-based arrangements involving actors outside organisational boundaries. It demonstrates the centrality of data as a resource for data-based organising and of data structures as devices for governing the data collection process.

The contribution by the thesis of a unique perspective for the understanding of social media is highlighted through this contrast between the social media network *PatientsLikeMe* and the mobile operator. First, the contrast shows that these arrangements are characterized by some continuities, namely, the need to govern user bases towards specific goals of data generation and collection, and the concern with the

intrinsically uncertain enterprise of turning data into information which is valuable to businesses. Data are shown to be structures with specific properties and powers. They interact with local and semantic contexts differently, depending on their structural configurations, and the actor's ability to make sense of them. Second, the thesis also shows that social media significantly differ from other types of open, distributed and data-based networks (see also Jonsson et al. 2009). They differ for the interpenetration of information production processes with sociality, and the instrumental construction of the latter that this relationship entails. For the accomplishment of these goals, complex data structures and comprehensive data collection processes are necessary, which need to be continually refined and tuned.

I believe that no fully exhaustive account of social media is possible, as perhaps with any other social phenomena of some relevance. But I would like to conclude the thesis by arguing that through the empirical evidence and its explanation as provided throughout the paper portfolio, I have provided answers covering the propaedeutic research questions I initially formulated in the introduction. To start with, the first question asked *how does a social media organisation develop technology in order to govern the platform's user base?* We have seen that a social media organisation develops technology under particular conditions. I characterized this situation by elaborating the perspective on information production, which connects the organisation's business model (and instrumental relationship with the user base) with the specific technology development activities that the organisation undertakes. I have started to discuss the methodological circumstances of the perspective since the first paper, where I have argued for the case of CR in connection with the Batesonian event-theory of information. In the first and second paper the conditions for the emergence of

information from the collected data have also been fleshed out while the difference between the two case study organisations has emerged in the specific data management practices of the social media organisation.

The second question asked *how does social media technology support the management of the network?* This question, immediately following from the first, has been answered through an argument for the criticality of the data and data structures management practices as the main device through which an open, distributed and data-based organisational form can be orchestrated. At the centre of the three papers on the social media case stand a set of complex data architectures, advanced data aggregation and representation techniques and web-based technological solutions supporting easy and continuous updatability of data and code-base. This set of technological solutions is shown to be linked to the eventual success of the social media organisation in achieving the intended level of orchestration of the patient contribution, data collection inflow and consequent information production.

Then, the third question asked *how does a social media organisation manage its information production process?* I have characterised the set of processes and techniques of management of a social media network through a set of novel concepts, each connecting to specific conditions or circumstances of the information production perspective on social media. These are the concepts of *information cultivation* strategy, the associated mechanisms of *data pool enrichment* and *data pool extension*, and the technique of *social denomination.* I was then able to associate the information production process with the piecemeal fragmentation of data collection tasks, the development of solutions that guide users of different (medical) literacy through input,

and most importantly the engagement and motivation of (patient) users operating through a complex set of interdependencies between data generation and self-representation and socialization opportunities.

Finally, the fourth question asked *how are social media networks shaping sociality and the users' life context?* I have started to answer to this broad and paramount question first through the development of the concept of *social denomination,* showing in detail how the techniques through which a social media infrastructure is managed for the production of information continuously reshape the social environment the users are supposed to adapt to. Explaining the context of the new *computed sociality* I have shown how seemingly inert data representations are instead involved in the dynamic construction and reconstruction of the social media environment. Social media are spaces engineered to draw select interaction trajectories, for eliciting desired levels of engagement and user activity. The consequences of this instrumentalization of sociality are to be found, I argue, in the shaping of individual self-representation, social and experiential life context, by favouring certain metaphors of the patient experience, and highlighting specific treatment options or lifestyles rather than others, and so on. Also further consequences of an instrumentalization of sociality that aims at optimizing its productivity is the proliferation of a whole range of new (medical) objects, created through simple system interactions and immediately legitimized by their association with numbers and scores that give them social dimension and weight.

In sum, the thesis is not limited to discussing the engineering of sociality, but it does so in a framework that centres on information production processes from an organisational perspective, and contributes an essential theorisation of the critical

structures involved, the most prominent of which are data structures. As such, I have demonstrated how the implications are broad and profound. Information production through social media provides an alternative to established data collection and management practices traditionally involving institutions and professions, in key domains of society. It also reconfigures relationships between extant forms of knowledge and epistemic evidence.

Perhaps, when confronted with all the issues that one can immediately relate to social media, the thesis might appear narrowly focused. All of us have first-hand experience of social media, and as experts in a discipline at the crossroads of social and technical domains, we probably have strong opinions about the most significant character and implications of these developments. There are very important questions of broad public interest, about social media. One might be tempted to seek an answer to whether these social media systems really have the potential of realising, in a parcelled and diffused way, a more democratic society, where the construction of facts and truth, and execution of rules and assignment of roles, are remit to social groups larger and more inclusive than before (e.g. Bowker, 2013; Callon, 2009; Feenberg, 2010). In addition, there are important data access, ownership and other intellectual property questions that it would be important to answer (e.g. Lessig, 2008; Lunschof *et al.,* 2014). This connects to other concerns as to whether the capillary and systematic end user involvement that these systems afford constitutes labour of a status that should correspond to some form of remuneration. Clearly users obtain value through improved level of service in exchange for basically every data sharing activity – and this was true for "offline" services too – but this does not automatically counter arguments that call for a more even redistribution of the riches that these networks (sometimes) are able to

generate. In the era of Big Data, user data is an asset which at the same time is difficult to evaluate (Mayer-Schönberger and Cukier, 2013; Tempini, 2013) and also critically underpins a few towering empires (implicitly, throughout the thesis it is demonstrated why such valuation is difficult to realise: information is not fully contained in the data, instead, it crucially depends on an organisation's skills and context). These are just few of the many questions left open. Still, I believe that the thesis' set of papers identifies a set of topics that are essential to any in-depth understanding of the social media phenomena at least as the first 10 years of their existence are concerneed. In fact, concerning social media we can literally say 'information is the currency of the age' (Kallinikos, 2011:72), bringing together unforeseen configurations of social actors. Putting data (generation and collection) and information (cultivation and actualisation) at the centre of analyses of social media networks should perhaps be a central feature for many solid investigations of social media phenomena.

# References


Aaltonen, A. (2012) The Beauty and Perils of Metrics, *Mercury Magazine* 1(3): 56-59.

Aaltonen, A. and Kallinikos, J. (2013) Coordination and Learning in Wikipedia: Revisiting the Dynamics of Exploitation and Exploration, In M. Holmqvist & A. Spicer (Eds.), *Research in the Sociology of Organisations* , Vol. 37, Emerald Group Publishing Limited, pp. 161–192.

Aaltonen, A. and Tempini, N. (2014) Everything counts in large amounts: a critical realist case study on data-based production, *Journal of Information Technology* 29(1): 97–110.

Abbott, A. (1988) Things of boundaries, *Social Research* 62(4): 857–882.

Abbott, A. (2001) *Time matters: On Theory and Method,* Chicago: The University of Chicago Press.

Agamben, G. (2009) *What is an Apparatus? And Other Essays*, Stanford: Stanford University Press.

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

Alaimo, C. (2014) *Computational Consumption: Social media and the construction of digital consumers,* PhD Thesis. Department of Management, London School of Economics.

Archer, M. S. (1982) Morphogenesis Versus Structuration: On Combining Structure and Action, *The British Journal of Sociology* 33(4): 455-483.

Archer, M. S. (1998) Introduction: Realism in the Social Sciences, In M. Archer, R. Bhaskar, A. Collier, T. Lawson and A. Norrie, (eds.) *Critical Realism: Essential Readings*, New York: Routledge, pp. 189-205.

Arnott-Smith, C. and Wicks, P. (2008) PatientsLikeMe: Consumer Health Vocabulary as a Folksonomy, *AMIA Annual Symposium Proceedings* 2008: 682–686.

Ascher, S. and Jordan, J. (2006) *So Much, So Fast*, West City Films.

Avgerou, C. (2013) Social Mechanisms for Causal Explanation in Social Theory Based IS Research, *Journal of the Association for Information Systems* 14(8): 399–419.

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

Baskerville, R. L. and Myers, M. D. (2002) Information Systems as a Reference Discipline, *MIS Quarterly* 26(1): 1-14.

Becker, H. S. (2007) *Writing for Social Scientists: How to Start and Finish Your Thesis, Book, or Article*, Chicago: University of Chicago Press.

Beniger, J. R. (1986) *The control revolution: Technological and economic origins of the information society*, Cambridge, MA: Harvard University Press.

Benkler, Y. (2007) *The wealth of networks: How social production transforms markets and freedom*, New Haven & London: Yale University Press.

Berg, M. (1997) Of Forms, Containers, and the Electronic Medical Record: Some Tools for a Sociology of the Formal. *Science, Technology, & Human Values* 22(4): 403–433.

Berg, M. (2004) *Health Information Management,* London: Routledge.

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

Benbasat, I. and Zmud, R. W. (2003) The Identity Crisis within the IS Discipline: Defining and Communicating the Discipline's Core Properties, *MIS Quarterly* 27(2): 183-194.

Benkler, Y. (2006) *The Wealth of Networks: How Social Production Transforms Markets and Freedom*, New Haven: Yale University Press.

Bermejo, F. (2009) Audience Manufacture in Historical Perspective: From Broadcasting to Google, *New Media & Society* 11(1/2): 133-154.

Bhaskar, R. (1998) Philosophy and Scientific Realism, In M. Archer, R. Bhaskar, A. Collier, T. Lawson and A. Norrie, (eds.) *Critical Realism: Essential Readings*, New York: Routledge.

Bhaskar, R. (2008) *A Realist Theory of Science*, New York: Routledge.

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). Retrieved from http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/3037/2568

Bowker, G. C. (2005) *Memory Practices in the Sciences*, Cambridge, MA: MIT Press.

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

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

boyd, d. m. and Crawford, K. (2012) Critical Questions for Big Data. Provocations for a cultural, technological, and scholarly phenomenon., *Information, Communication & Society* 15(5): 662–679.

boyd, d. m. and Ellison, N. B. (2008) Social Network Sites: Definition, History, and Scholarship, *Journal of Computer-Mediated Communication* 13(1): 210–230.

Bratich, J. Z. (2005) Amassing the Multitude: Revisiting Early Audience Studies, *Communication Theory* 15(3): 242-265.

Brewer, J. D. (2000) *Ethnography*, Buckingham: Open University Press.

Bygstad, B. (2010) Generative Mechanisms for Innovation in Information Infrastructures, *Information and Organisation* 20(3-4): 156-168.

Callon, M. (2009) *Acting in an uncertain world : an essay on technical democracy*, Cambridge, Mass: MIT Press.

Canguilhem, G. (2012) *Writings on Medicine*, New York, NY: Fordham University Press.

Carr, N. G. (2008) *The Big Switch*, New York: W. W. Norton & Company.

Cheney-Lippold, J. (2011) A New Algorithmic Identity: Soft Biopolitics and the Modulation of Control, *Theory, Culture & Society* 28(6): 164–181.

Ciborra, C. (2002) *The labyrinths of information: challenging the wisdom of systems.*, Oxford: Oxford University Press.

Ciborra, C. (2003) Hospitality and IT, In F. Ljundberg (Ed.), *Informatics in the Next Millennium,* Lund: Studentlitteratur, pp. 161–173.

Clarke, A. E., Shim, J. K., Mamo, L., Fosket, J. R. and Fishman, J. R. (2003) Biomedicalization: Technoscientific transformations of health, illness, and US biomedicine, *American sociological review* 68(2): 161–194.

Cohen, S. M. (2012) Aristotle's Metaphysics, In (E. N. Zalta, Ed.) *The Stanford Encyclopedia of Philosophy*. Retrieved from http://plato.stanford.edu/archives/sum2012/entries/aristotle-metaphysics/

Conrad, P. (2005) The Shifting Engines of Medicalization. *Journal of Health and Social Behaviour* 46(1) 3–14.

Corbin, J. and Strauss, A. (2008) *Basics of qualitative research: Techniques and procedures for developing grounded theory*, London: Sage.

Crotty, M. (1998) *The Foundations of Social Research: Meaning and Perspective in the Research Process*. London: SAGE Publications.

Day, G. S. (2003) Creating a Superior Customer-relating Capability, *MIT Sloan Management Review* 44(3): 77-82.

Desrosières, A. (1998) *The Politics of Large Numbers: A History of Statistical Reasoning*, Cambridge, MA: Harvard University Press.

Dodier, N. (1998) Clinical Practices and Procedures in Occupational Medicine: A Study of the Framing of Individuals, In M. Berg, A. Mol, (eds.), *Differences in Medicine: Unraveling Practices, Techniques, and Bodies,* London: Duke University Press, pp. 53–85.

Dourish, P. (2001) *Where the Action Is*, London: The MIT Press.

Dreyfus, H. L., S. E. Dreyfus. (1986) *Mind Over Machine*, New York: Free Press.

Easton, G. (2010) Critical realism in case study research, *Industrial Marketing Management* 39: 118-128.

Ekbia, H. R. (2009) Digital Artifacts as Quasi-Objects: Qualification, Mediation, and Materiality, *Journal of the American Society for Information Science and Technology* 60(12): 2554-2566.

Elder-Vass, D. (2005) Emergence and the Realist Account of Cause, *Journal of Critical Realism* 4(2): 315-338.

Elder-Vass, D. (2007) For Emergence: Refining Archer's Account of Social Structure, *Journal for the Theory of Social Behaviour* 37(1): 25-44.

Epstein, S. (1996) *Impure Science: AIDS, Activism, and the Politics of Knowledge*, London: University of California Press.

Epstein, S. (2008) Patient Groups and Health Movements, In E. J. Hackett, O. Amsterdamska, M. Lynch, & J. Wajcman (Eds.), *The Handbook of Science and Technology Studies*, London: MIT Press, pp. 499–539.

Ettema, J. S. and Whitney, D. C. (1994) The Money Arrow: An Introduction to Audiencemaking, in Audiencemaking: How the Media Create the Audience, In J. S.

Ettema and D. C. Whitney, (eds.) *Sage Annual Reviews of Communication Research*, London: SAGE Publications, pp. 1-18.

Faraj, S., Jarvenpaa, S. L. and Majchrzak, A. (2011) Knowledge Collaboration in Online Communities, *Organisation Science* 22(5): 1224–1239.

Faulkner, P. and Runde, J. (2009) On the Identity of Technological Objects and User Innovations in Function, *Academy of Management Review* 34(3): 442–462.

Faulkner, P. and Runde, J. (2010) The Social, the Material, and the Ontology of Non-Material Objects, In Judge Us Seminar (University of Cambridge, UK, 2010).

Faulkner, P. and Runde, J. (2013) Technological Objects, Social Positions, and the Transformational Model of Social Activity, *MIS Quarterly* 37(3): 803-818.

Feenberg, A. (2010) *Between Reason and Experience: Essays in Technology and Modernity*, Cambridge, MA: MIT Press.

Feenberg, A. L., Licht, J. M., Kane, K. P., Moran, K. and Smith, R. A. (1996) The online patient meeting, *Journal of the Neurological Sciences* 139, Supplement: 129–131.

Flick, U. (2004) Triangulation in Qualitative Research, In U. Flick, E. von Kardoff and I. Steinke, (eds.) *A Companion to Qualitative Research,* London: SAGE Publications, pp. 178-183.

Flyvbjerg, B. (2006) Five Misunderstandings About Case-Study Research. *Qualitative Inquiry* 12(2) 219-245.

Gerlitz, C. and Helmond, A. (2013) The Like economy: Social buttons and the data-intensive web, *New Media & Society* 15(8): 1348–1365.

Gieryn, T. F. (1983) Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists, *American Sociological Review* 48(6): 781–795.

Hacking, I. (1990) *The Taming of Chance*. Cambridge: Cambridge University Press.

Hacking, I. (1999) *The Social Construction of What?* Cambridge, MA: Harvard University Press.

Hanseth, O. (1996) *Information Technology as Infrastructure.* Doctoral Dissertation. Göteborg University.

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

Hasselbladh, H. and Bejerot, E. (2007) Webs of Knowledge and Circuits of Communication: Constructing Rationalized Agency in Swedish Health Care. *Organisation* 14(2) 175–200.

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.

Heidegger, M. (1977) *The Question Concerning Technology and Other Essays*, New York: Harper and Row.

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.

Jonsson, K., Holmström, J. and Lyytinen, K. (2009) Turn to the material: Remote diagnostics systems and new forms of boundary-spanning. *Information and Organisation* 19(4) 233–252.

Kallinikos, J. (1993) Identity, recursiveness and change: Semiotics and beyond, In P. Ahonen (Ed.), *Tracing the Semiotic Boundaries of Politics*, Berlin: Mouton de Gruyter, pp. 257–278.

Kallinikos, J. (1995) The Architecture of the Invisible: Technology is Representation. *Organisation* 2(1) 117–140.

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. (2009) On the Computational Rendition of Reality : Artefacts and Human Agency, *Organisation* 16(2): 182–202.

Kallinikos, J. (2010) Smart Machines. *Encyclopedia of Software Engineering,* London: Taylor and Francis, pp. 1097–1103.

Kallinikos, J. (2011) *Governing Through Technology: Information Artefacts and Social Practice*, Basingstoke: Palgrave Macmillan.

Kallinikos, J. (2012) The Allure of Big Data, *ParisTech Review*. Retrieved from http://www.paristechreview.com/2012/11/16/allure-big-data

Kallinikos, J., Aaltonen, A. V. and Marton, A. (2013) The Ambivalent Ontology of Digital Artifacts, *MIS Quarterly* 37(2): 357–370.

Kallinikos, J. and Tempini, N. (2014) Patient Data as Medical Facts: Social Media Practices as a Foundation for Medical Knowledge Creation, *Information Systems Research* 25(4): 817-833.

Kallinikos, J. and Tempini, N. (2011) Post-material Meditations: On Data Tokens, Knowledge and Behaviour, Presented at the 27th EGOS Colloquium - European Group of Organisational Studies, Gothenburg. Retrieved from http://www.tigair.info/docs/kalltemp_egos11.pdf

Kelly, K. (1996) The Electronic Hive: Embrace It, In R. Kling (Ed.), *Computerization and Controversy: Value Conflicts and Social Choices*, 2nd ed., San Francisco, CA: Morgan Kaufmann Publishers, pp. 75–78.

Langley, A. (1999) Strategies for theorizing from process data, *Academy of Management Review* 24(4): 691-710.

Latour, B. (1987) *Science in Action*, Cambridge, MA: Harvard University Press.

Latour, B. (1999) *Pandora's Hope: Essays on the Reality of Science Studies*, Cambridge, MA: Harvard University Press.

Latour, B. (2000) When things strike back: a possible contribution of 'science studies' to the social sciences, *British Journal of Sociology* 51(1): 107–123.

Lee, A. S. (2010) Retrospect and Prospect: Information Systems Research in the Last and Next Twenty-Five Years, *Journal of Information Technology* 25(4): 336–348.

Leonardi, P. M. (2010). Digital Materiality? How Artifacts Without Matter, Matter, *First Monday* 15(6): 1–15. Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/3036/2567

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

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.

Lessig, L. (2008) *Remix: Making Art and Commerce Thrive in the Hybrid Economy*, London: Bloomsbury Academic.

Lyytinen, K. and Yoo, Y. (2002) The Next Wave of Nomadic Computing, *Information Systems Research* 13(4): 377-388.

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

Lunshof, J. E., Church, G. M. and Prainsack, B. (2014) Raw Personal Data: Providing Access, *Science* 343(6169): 373–374.

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

Marks, H. M. (1997) *The progress of experiment. Science and therapeutic reform in the United States, 1900-1990*, Cambridge: Cambridge University Press.

Markus, M. L. (1983) Power, Politics, and MIS Implementation, *Communications of the ACM* 26(6): 430–444.

Mathiassen, L. and Stage, J. (1990) The principle of limited reduction in software design, *Information Technology & People* 6(2/3): 171–185.

Mathiassen, L. and Sorensen, C. (2008) Towards a theory of organisational information services, *Journal of Information Technology* 23(4): 313–329.

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

McKinney, E. H. and Yoos, C. J. (2010) Information about Information: A Taxonomy of Views, *MIS Quarterly* 34(2): 329-344.

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

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

Mingers, J. and Willcocks, L. (2014) An integrative semiotic framework for information systems: The social, personal and material worlds, *Information and Organisation* 24(1): 48–70.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M. and Altman, D. G. (2010) CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ (Clinical research ed.)* 340 c869.

Morley, D. (2006) Unanswered Questions in Audience Research, *The Communication Review* 9(2): 101-121.

Morris, J. W. (2012) Making music behave: Metadata and the digital music commodity, *New Media & Society* 14(5): 850–866.

Morville, P. (2005) *Ambient Findability*, Sebastopol, CA: O'Reilly Media.

Mutch, A. (2010) Technology, Organisation and Structure – A Morphogenetic Approach, *Organisation Science* 21(2): 507-520.

Napoli, P. M. (2003) *Audience Economics: Media Institutions and the Audience Marketplace*, New York: Columbia University Press.

Orlikowski, W. J. (1996) Improvising organisational transformation over time: a situated change perspective, *Information Systems Research* 7(1): 63–92.

Orlikowski, W. J. (2007) Sociomaterial Practices: Exploring Technology at Work, *Organisation Studies* 28(9): 1435-1448.

Orlikowski, W. J. and Barley S. R. (2001) Technology and Institutions: What Can Research on Information Technology and Research on Organisations Learn from Each Other? *MIS Quarterly* 25(2): 145–165.

Pollock, N. and Williams, R. (2009) *Software and Organisations: the biography of the enterprise-wide system or how SAP conquered the world*, London: Routledge.

Porter, T. M. (1995) *Trust In Numbers : The Pursuit of Objectivity In Science and Public Life*, Princeton: Princeton University Press.

Prainsack, B. (2014) Beyond Professional Expertise. *Unpublished Manuscript: King's College, University of London.*

Rabeharisoa, V., Moreira, T. and Akrich, M. (2013) *Evidence-based activism: Patients' organisations, users' and activist's groups in knowledge society* (Working Paper No. 033), Paris, France: Centre de Sociologie de l'Innovation, Mines ParisTech.

Rajan, K. S. and Leonelli, S. (2013) Introduction: Biomedical Trans-actions, Postgenomics, and Knowledge/Value, *Public Culture* 25(3 71): 463–475.

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

Redman, T. C. (2008) *Data Driven*, Boston: Harvard Business Press.

Ribes, D. and Bowker, G. C. (2009) Between meaning and machine: Learning to represent the knowledge of communities, *Information and Organisation* 19(4): 199–217.

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

Rose, N. (1991) Governing by numbers: Figuring out democracy, *Accounting, Organisations and Society* 16(7): 673–692.

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

Rose, N. (2006) Disorders Without Borders? The Expanding Scope of Psychiatric Practice, *BioSocieties* 1(04): 465–484.

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

Rose, N. (2009) Normality and pathology in a biomedical age, *Sociological Review* 57: 66–83.

Rosenberg, D. (2013) Data before the Fact, In L. Gitelman (Ed.), *'Raw Data' Is an Oxymoron*, Cambridge, MA: MIT Press, pp. 15–40.

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

Runde, J., Jones, M., Munir, K. and Nikolychuk, L. (2009) On Technological Objects and the Adoption of Technological Product Innovations: Rules, Routines and the Transition From Analogue Photography to Digital Imaging, *Cambridge Journal of Economics* 33(1): 1-24.

Russell, B. (1994) *History of Western Philosophy*, Routledge: London.

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

Scheuermann, R. H., Ceusters, W. and Smith, B. (2009) Toward an Ontological Treatment of Disease and Diagnosis, In *Summit on Translational Bioinformatics*, Vol. 2009, pp.

116–120. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041577/

Schulz, K. F., Altman, D. G. and Moher, D. (2010) CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials, *BMJ (Clinical research ed.)* 340: c332.

Scott, S. V. and Orlikowski, W. J. (2012) Great Expectations: The Materiality of Commensurability in Social Media, In P. M. Leonardi, B. A. Nardi, & J. Kallinikos (Eds.), *Materiality and Organising: Social Interaction in a Technological World*, Oxford: Oxford University Press, pp. 113–133.

Shirky, C. (2008) *Here Comes Everybody: The Power of Organising Without Organisations*, London: Penguin Press.

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

Shryock, R. H. (1961) The History of Quantification in Medical Science, *Isis* 52(2): 215–237.

Simon, H. A. (1996) *The Sciences of the Artificial*, Cambridge, MA: The MIT Press.

Sismondo, S. (1993) Some Social Constructions, *Social Studies of Science* 23(3): 515-553.

Smith, G. (2008) *Tagging. People-Powered Metadata for the Social Web*, Berkeley, CA: New Riders.

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

Tempini, N. (2013) Book review: big data: a revolution that will transform how we live, work, and think, *LSE Review of Books*. Retrieved from http://blogs.lse.ac.uk/lsereviewofbooks/2013/05/02/book-review-big-data-a-revolution-that-will-transform-how-we-live-work-and-think/

Tempini, N. (in press) Governing PatientsLikeMe: information production and research through an open, distributed and data-based social media network, *The Information Society*.

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

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

Timmermans, S. and Oh, H. (2010) The Continued Social Transformation of the Medical Profession, *Journal of Health and Social Behavior* 51(S): S94–S106.

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

Tousijn, W. (2002) Medical dominance in Italy: a partial decline, *Social Science & Medicine* 55(5): 733–741.

Treem, J. W. and Leonardi, P. M. (2012) Social Media Use in Organisations: Exploring the Affordances of Visibility, Editability, Persistence, and Association, *Communication Yearbook* 36: 143–189.

Turner, M. R., Wicks, P., Brownstein, C. A., Massagli, M. P., Toronjo, M., Talbot, K. and Al-Chalabi, A. (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, NY: Oxford University Press.

Van Maanen, J. (1979) The Fact of Fiction in Organisational Ethnography, *Administrative Science Quarterly* 24(4): 539–550.

Van Maanen, J. (1993) *Tales of the Field*, Cambridge, MA: MIT Press.

Walsh, D. (1998) Doing Ethnography, in C. Seale, (ed.) *Researching society and culture*, London: SAGE Publications, pp. 217-232.

Weick, K. E. (1995) What Theory Is Not, Theorizing Is, *Administrative Science Quarterly* 40(3): 385-390

Weinberger, D. (2007) *Everything is Miscellaneous: The Power of the New Digital Disorder*, New York: Times Books.

Weiner, J. (2004) *His Brother's Keeper: A Story from the Edge of Medicine*, New York, NY: Harper Collins.

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

Wicks, P. and Frost, J. (2008) ALS patients request more information about cognitive symptoms, *European Journal of Neurology* 15(5): 497–500.

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

Wicks, P., Massagli, M. P., Frost, J., Brownstein, C., Okun, S., Vaughan, T. E., Bradley, R. and Heywood, J. (2010) Sharing Health Data for Better Outcomes on PatientsLikeMe, *Journal of Medical Internet Research* 12(2): e19.

Wicks, P., Vaughan, T. E., Massagli, M. P. and Heywood, J. (2011) Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm, *Nature Biotechnology* 29(5): 411–414.

Wikipedia. (2014) Evidence of absence, *Wikipedia, the Free Encyclopedia*. Retrieved May 5, 2014, from http://en.wikipedia.org/w/index.php?title=Evidence_of_absence&oldid=606482435

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

Wittgenstein, L. (1953) *Philosophical Investigations*, Oxford: Blackwell.

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

Wynne, B. (1996) May the Sheep Safely Graze? A Reflexive View of the Expert-Lay Knowledge Divide, In S. Lash, B. Szerszynski, & B. Wynne (Eds.), *Risk, Environment and Modernity. Towards a New Ecology*, London: Sage, pp. 44–83.

Yates, J. A. (1989) *Control Through Communication: The Rise of System in American Management*, Baltimore: Johns Hopkins University Press.

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

Yoo, Y. (2010) Computing in Everyday Life: A Call for Research on Experiential Computing, *MIS Quarterly* 34(2): 213-231.

Yoo, Y., Boland, R. J. Jr., Lyytinen, K. and Majchrzak, A. (2012) Organising for Innovation in the Digitized World, *Organisation Science* 23(5): 1398-1408.

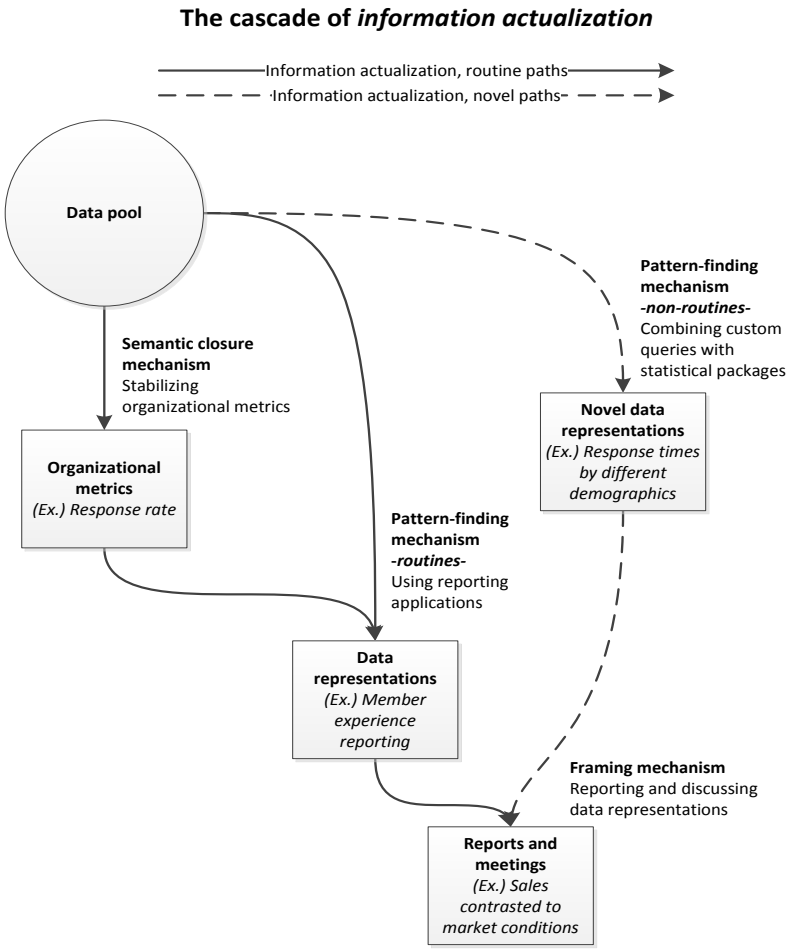
Yoo, Y., Henfridsson, O. and Lyytinen, K. (2010) The New Organising Logic of Digital Innovation: An Agenda for Information Systems Research, *Information Systems Research* 21(4): 724-735.

Zittrain, J. L. (2008) *The Future of the Internet and How to Stop It*, New Haven: Yale University Press.

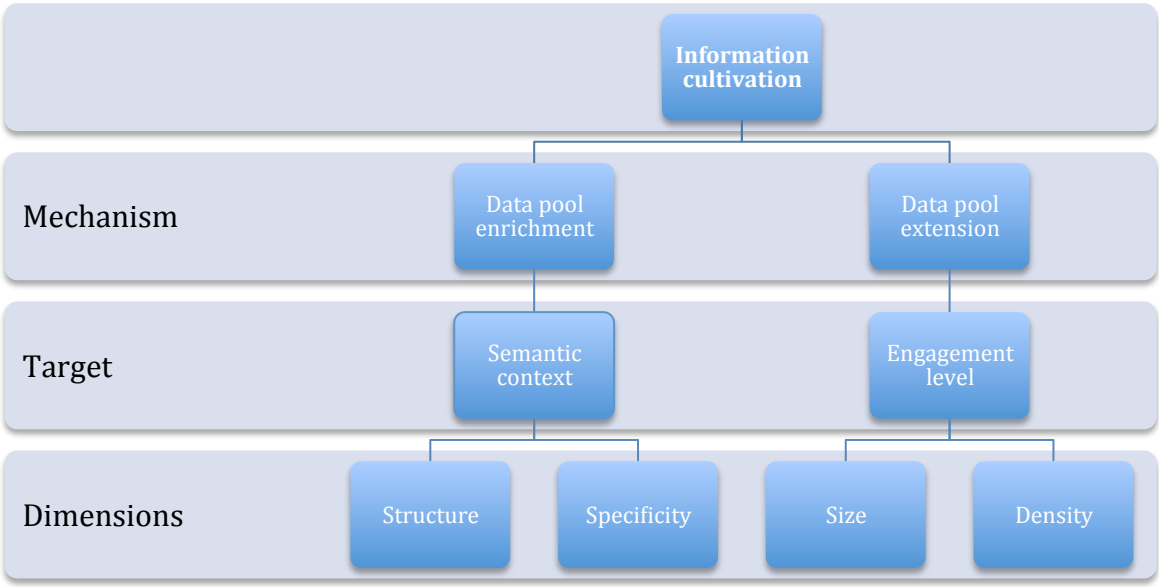
Zuboff, S. (1988) *In the Age of the Smart Machine: The Future of Work and Power*, New York, NY: Basic Books.

# Appendix

## Appendix 1



## Appendix 2


Information cultivation
Mechanism
Data pool enrichment
Data pool extension
Target
Semantic context
Engagement level
Dimensions
Structure
Specificity
Size
Density