BUILDING SOCIAL MEDIA THEORY FROM CASE STUDIES:

A NEW FRONTIER FOR IS RESEARCH

Abstract

This paper concentrates on a major concern in the IS field – theory building – and couples it with a major development in our field, social media-related environments, to consider how we might build theory, using digital texts, within the case study methodology.

The growing popularity and constant innovations of social media platforms and applications have transformed ways of interacting, working, creating value and innovating. There is a need – to theorize these new environments, and the intriguing social and technical dynamics they make possible. We elaborate upon how building theory from case studies should be adapted to the opportunities and challenges of social media environments. We delve into key challenges of the research process: case study design, data analysis, and engaging in multi methods. Doing so, we identify some key considerations that can help IS researchers navigate the still new and not yet fully understood characteristics of these environments for theory building purposes.

Keywords: Theory building, methods, social media, case studies, digital text, mixed methods.
Introduction

In the last few years, new generations of web applications have dramatically increased people’s ability to interact with each other electronically, to generate content online, and to immerse themselves in alternative universes (Boyd and Ellison 2008; Damer 2008; Messinger et al. 2009; O’Reilly 2007). These social media have recently become widespread, and led to new social practices (Koebler et al. 2010), sociability patterns (Van Den Eede 2010), learning practices (Greenhow and Robelia 2009; Kim and Abbas 2010), leisure activities, as well as social and political mobilization (Byrne 2008; Wattal et al. 2010). Social media use has also permeated the business domain (Culnan et al. 2010), triggering new business models (Lyons 2008), customer-relationship tactics (Di Gangi et al. 2010; Gallaugher and Ransbotham 2010) and managerial practices (Kaganer and Vaast 2010; Leidner et al. 2010).

The potential for social media research is huge, and there is a strong need for researchers to try and make sense of these new dynamics. Studies in sociology, marketing and Information Systems (IS), have all examined social media in some depth (e.g. Beer and Burrows 2007; Berente & al. 2011; Cooke and Buckley 2009). Yet, much of this research has been descriptive, i.e. it has presented empirically, in sometimes great detail, what these new environments are. There is therefore a growing opportunity for IS researchers to move beyond describing and towards theorizing these new contexts and associated practices (Majchrzak 2009).

The general and critical concern expressed in the last ten years or so about the lack of theory in the IS discipline as a whole (Watson 2001; Weber 2003) arguably still applies, and is especially apparent in social media research. Theory can be simply defined as “a statement of concepts and their interrelationships that shows how and / or why a phenomenon occurs” (Corley and Gioia 2011, p. 12). Gregor (2006) gives a useful typology of IS theories: theory for analyzing, theory for explanation, theory for prediction, theory for both explanation and prediction, and finally, theory for design and action. Social media-related environments are promising venues to develop these different types of theories. Firstly, for analyzing e.g. how are crises relayed by and partly dealt with social media (Majchrzak 2011)?. Secondly, for explaining e.g how does membership turnover pattern affect knowledge creation in wiki-based environments (Ransbotham & Kane, 2011)?. Finally for predicting e.g. which user-generated video will go viral (Nahon et al. 2011)?. We are interested in theory building that explicitly considers relationships between concepts as opposed to detailed description. This may, or may not, in our view, encompass causal relationships, and this is something we discuss later with regard to what theory building potential exists with ‘digital texts,’ which are generally exclusively qualitative texts. So one additional consideration in our paper is the use of mixed methods for theory building. We also consider how theory building tools inherent in grounded theory methodology (such as theoretical sampling to expand the scope of a theory) are currently underused in the information systems discipline (Urquhart 2012, Forthcoming). This paper therefore deals with a two-fold concern about building social media theory. Firstly, it deals with the need to build theory about these social media-related environments, as called forth by Kane et al. (2012). Secondly, it deals with a related need that has so-far not fully been examined, that of considering how to build theory in these new contexts.

This paper thus explores the theory building potential of case studies of social media, and offers some methodological considerations for IS researchers.

We focus on theory building from case studies in social media-based environments, and highlight the opportunities and challenges that IS researchers are likely to face in this pursuit. It seems urgent for the IS community to develop ways of building theory for social media because many IS researchers have embraced these environments as contexts for their research (e.g. Ransbotham and Kane 2011; Wattal et al. 2010), and many more have been thinking about doing so. One can only anticipate that many more future publications will deal with social media environments. Consequently, there has been a growing concern, among social media IS scholars (e.g. Kane and Fichman 2009; Kane et al. 2012; Majchrzak 2009; Te’eni 2009) that the IS community has not yet been sufficiently engaged in reflecting upon the methodological aspects of researching social media, and subsequent implications for theory building. In particular, there has not yet been much discussion of what it takes to build theories of these new environments. This paper aims to elaborate on these emerging discussions in the IS community by considering the theory building process in terms of the data that the researcher can collect from social
media. We also examine new and possibly unexpected methodological dilemmas the researcher may face, and suggest ways to collect and analyze data from these environments in order to build theory. It is not our ambition to provide the last word on these important issues. This would be unrealistic, especially as social media and the intricately related social and technical conditions they generate are in flux, with a stream of new applications becoming available and adopted. Rather, building upon our own and other scholars’ (e.g. Fleming and Waguespack 2007) experience researching these contexts, we present documented arguments for taking theory building seriously when investigating social media. It is our belief that the IS community has much to benefit by considering such issues, and that theory building will help us to stay important and relevant as an academic field.

Our paper is organized as follows. We start by highlighting the importance for IS research of theorizing social media. We then examine in more detail the use of case studies for theory building, and what might constitute a digital text in social media. We then discuss three challenges and key methodological considerations for IS researchers – Case Study Design, Data Analysis, and Engaging in Mixed Methods. We conclude by highlighting some of the main implications for IS researchers of engaging with social media to build theory.

Building Social Media Theory: A New Frontier for IS Research

Setting the stage for our paper we start by presenting what we mean by “social media environments,” and by arguing both that IS researchers would benefit from tackling the theorizing of social media and that they would be especially well-suited to do so.

Social media

Social media, or “Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content” (Kaplan and Haenlein 2010, p. 62), have generated unprecedented opportunities for the development of new theories in IS research, as well as unanticipated methodological challenges for researchers. Opportunities lie in the multiplicity and diversity of methodological approaches that researchers might adopt in these environments: e.g. experiments (Antheunis and Scouten 2010; Minocha et al. 2010), ethnography (Garcia et al. 2009; Kien 2008), “cyber-archeology” (Zimbra et al. 2010), or case studies (Vaast et al. Forthcoming; Veer 2011). The methodological challenges of dealing with these new environments are however also varied, and range from the technical (how to do it?) (Bollier 2010; Boyd and Ellison 2008), to the legal (how legally accountable is the researcher?) (Allen et al. 2006; Lehmberg et al. 2008), and the ethical (what are the right courses of action?) (Beer 2008; Stanton 2010).

Social media environments vary substantially, as the table below illustrates (Table 1).

<table>
<thead>
<tr>
<th>Social media applications and environments</th>
<th>Examples</th>
<th>Examples of publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social networking sites</td>
<td>Facebook, Myspace</td>
<td>(Grasmuck et al. 2009)</td>
</tr>
<tr>
<td>Wikis</td>
<td>Wikipedia</td>
<td>(Niederer and van Dijck 2010)</td>
</tr>
<tr>
<td>Blogging and microblogging</td>
<td>Huffington Post, Twitter</td>
<td>(Macias et al. 2009)</td>
</tr>
<tr>
<td>Virtual worlds</td>
<td>Second life</td>
<td>(Berente et al. 2011; Schultze and Orlikowski 2010)</td>
</tr>
<tr>
<td>Tagging, bookmarking, online reviews</td>
<td>Del.ic.ious</td>
<td>(Scott and Orlikowski 2010)</td>
</tr>
<tr>
<td></td>
<td>YouTube</td>
<td>(Arriaga and Levina 2010)</td>
</tr>
<tr>
<td>Photo and video sharing sites</td>
<td>YouTube</td>
<td>(Lange 2008)</td>
</tr>
</tbody>
</table>
This wide variety of environments would imply that there should also be a wide variety of methods (in particular qualitative and quantitative, automated or not, etc.) that can be appropriate to investigate such environments. We acknowledge this diversity, and advocate, in particular, that IS researchers consider the specifics of their environments. However, beyond this diversity of environments, there are common, and new, challenges and opportunities that deserve to be better taken in consideration.

As a caveat, we note here that it is not our intention to claim that everything about social media is new, and that the challenges and opportunities they afford IS researchers are always unprecedented. There is a body of research on online communities (e.g. Kraut et al. 1999; Ma and Agarwal 2007; Wilson and Peterson 2002) and open-source software development (e.g. Fleming and Waguespack 2007; O’Mahony and Ferraro 2007; Von Hippel and Von Krogh 2003) that has for some time illustrated some of these challenges and opportunities. Vaast and Walsham (Forthcoming), examined how grounded theorizing methods, in particular, might be adjusted in these computer-mediated and social environments. However, the fast growing and increasingly pervasive popularity of social media in many aspects of everyday life, and the highly dynamic character of social media applications (and of their waves of popularity) make it especially important for IS researchers to consider building theory of social media.

**Why the IS Discipline Should Tackle Social Media Theorizing**

One could reason that IS researchers do not have to develop theories of social media, because these environments might not be part of the “core” of our discipline. Such an argument, though, would be dubious. For one, some have disputed the very idea of a “core” of our discipline (Lyytinen and King 2006). Moreover, Lyytinen and King (2006) contend that academic legitimacy depends upon three drivers that justify the need for the IS community to engage in theorizing social media: the salience of issues, the maintenance of disciplinary plasticity, and the production of strong results.

**Salience of issues studied:** Developing social media theories is extremely important from a praxis perspective, because social media have changed the way we work, collaborate, innovate, manage, and organize. Therefore, it is important that theories that account for such crucial new developments be developed. For instance, knowledge creation and innovation processes have become more prevalent these last few years and have changed the way we need to think about them. New research on open innovation, for instance, has illustrated how salient these issues are, and how important it is to build fitting theories for them (Fleming and Waguespack 2007).

**Maintenance of disciplinary plasticity:** Lyytinen and King (2006) considered that “salience can be ephemeral, and for very good reasons” (p. 232). Robey (2003) also recommended “adaptive instability” for the IS discipline. Such advocated malleability can help keep the IS discipline remain salient in the future. In this regard, developing social media theories is important to keep the IS discipline abreast of changes taking place not only in business, but also in many other societal realms, including politics, education, etc. We note here, however, that we do not claim uncritically that anything social media-related is entirely and inherently new, and that completely new theories have to be built for everything social media-related. Some aspects and dynamics taking place with social media might be well explained by existing theoretical lenses, others require adjustment of existing theories, others still the development of new theories (Majchrzak 2009). What we argue, instead, is that, at the very least, there is a need for IS researchers to question their available theoretical “tool box” and adjust or expand it accordingly. In this regard, a growing group of IS researchers has become interested in studying issues that conflate the business with the societal and the political, especially because social media environments intricately relate these multiple dimensions (e.g. Vaast et al. 2012; Wattal et al. 2010)). Theory is needed to understand better how these issues arise and what behaviors they generate at multiple levels.

**Production of strong results:** Lyytinen and King (2006) considered that good theory is vital for the production of strong results. Extending this claim, this paper aims at guiding IS researchers in building social media theories that can bring about strong and important contributions to our field, and other disciplines. In this regard, we believe that theory plays an essential role in producing strong results and that IS researchers are well-placed to develop such theory. Indeed, many researchers from the computer science tradition have developed sophisticated algorithms to explore and mine high volume of social media data (e.g. Boyd et al. 2010; Boyd and Ellison 2008; Grace et al. 2010; Marwick and Boyd 2010) and have also pondered upon the methodological implications of social media research (Boyd and Crawford 2011). However, computer scientists have not yet much endeavored to theorize social media. In contrast,
the IS field benefits from a tradition of developing new theories and of building upon existing theories from other, related disciplines (Gregor, 2006; Majchrzak, 2011).

**IS Researchers are Especially Well-Suited to Develop such Theorizing**

Theory building of social media-related environments should combine technological and social aspects, and this dual consideration is a key characteristic that distinguishes IS research from research in other fields whose boundaries with the IS discipline are not always clear-cut (e.g., computer science, library sciences, sociology, general management). There has for instance been a recent and notable surge of interest for a “sociomaterial” perspective (e.g., Orlikowski 2007; Orlikowski and Scott 2008; Wagner et al. 2010) as well as for understanding social dynamics of practices and use associated with the features and affordances of systems (e.g., Fayard and Weeks 2007; Leonardi 2011; Zammuto et al. 2007) in the IS field. These exciting theoretical developments in the IS discipline are indicative of interest but also of the theoretical potential to open the “black box” of technology and to conceptualize the intricate relationships between the social and the material. To theorize dynamics and processes emerging from social media, such consideration is essential. We note, however, that theory of social media-related environments confronts researchers with an added, new and delicate, challenge. Researching and theorizing social media is difficult, because social media represent for researcher both sources of data and the sociomaterial context to be theorized (Boyd and Ellison 2008). To a certain extent, theorizing social media forces IS researchers to confront the “entanglement” (Orlikowski 2007; Scott and Orlikowski 2010) or the “imbrication” (Leonardi 2011; Leonardi and Barley 2008) of the material and the social into their own research practices.

IS researchers may be especially well-suited to tackle the theorization of social media-related environments, because they have already investigated and theorized highly dynamic and changing environments. For instance, Yoo et al.’s (2010) work on the “digitization” of innovation highlighted some of the implications for theorizing in the IS discipline of digital technologies characterized by constant innovations and layered changes. Building upon their insights, and applying them to the similarly and related dynamic context of social media, we consider that such characteristics have implications for theory building. In particular, IS researchers should have the ambition to develop mid-level, substantive theories of social media. Too broad or wide ranging theories might be “grander,” but they are likely to miss some of the intricacies and dynamics of these environments, especially as they relate to the imbrication between the social and the technological. Too narrow or focused, and the resulting theories are at risk of having a very short “shelf life,” constrained by the latest changes in technologies or by the relative popularity of social media platforms. Treem and Leonardi’s (2012) work on affordances of social media in organizational contexts presents an example of such interesting mid-range theory building of social media. On the basis of their analysis and synthesis of a wide range of studies in communication, HCI and system sciences disciplines, Treem and Leonardi (2012) identified visibility, persistence, editability, and association as key affordances of social media in organizations.

**Building Blocks for the Social Media Researcher**

In this section, we discuss what we see as two essential building blocks for the social media researcher interested in building theory from case studies. We first propose an understanding of theory building using case studies. We then provide an understanding of what the unit of analysis in a social-media related environment might be, through the notion of a digital text.

**Building Theory from Case Studies**

Torraco (2002) identified no less than five methods of theory building: 1) Dubin’s (1978) method for theory building, which takes a deductive approach using quantitative data; 2) grounded theory, where the theory has a close ties with the data; 3), statistical meta analysis of studies; 4) social constructionist theory building with an emphasis on local accounts of social worlds; and, finally, 5) theory building using case studies, where the aim is to produce explicit theoretical statements about the case study context.

We can see all these methods, to a greater or lesser degree, used in the IS discipline; we would also point out that grounded theory method has been frequently used in conjunction with case studies for the purposes of theory building (e.g., Levina and Vaast 2008; Orlikowski 1993).
Our paper concentrates on the use of case studies for building theory of social media, because the case study method affords researchers the ability to adjust to the specifics of the environment and gradually get to collect multiple types of data. The label “case study method” is an umbrella term, which means that developing a case study can (and should) rely upon multiple sources of evidence (e.g. documents, archives, records, participant observations, verbatim, etc.). This is especially well suited for the purpose of building theories of social media, because, as we will develop further below, these environments generate new data and create a need for creativity on researchers’ part in terms of data collection and analysis.

Eisenhardt (1989) is generally accepted to have pioneered the idea of using case studies for theory building in her seminal article, where she introduced an eight step model: Getting started, selecting cases, crafting instruments and protocols, entering the field, analyzing the data, shaping the hypotheses or theory, enfolding literature and reaching closure. The key consideration is that constructs should emerge from the case studies. In a similar way, but from an interpretive perspective, Walsham (1995) explained how analytical generalizations can come from case studies. He named four possible routes: developing concepts, generating theory, drawing specific implications for a particular domain that can be useful for similar concepts, and finally contributing to rich insight. What both seminal papers have in common are a dual emphasis on concept generation and engagement with existing theories. While Walsham’s (1995) paper does mention rich insight, the emphasis in both papers is on moving from description to analysis – exactly the same concern we have identified with social media studies. Eisenhardt and Graebner (2007), in reviewing the opportunities and challenges of theory building using case studies, point out that the theory building intent inherent in some case studies is often misunderstood, and say that it is a natural complement to mainstream deductive research. According to Eisenhardt and Graebner (2007), a key misunderstanding lies in thinking that the cases need to be representative, when in fact they are sampled for the purpose of developing theory. This process is known as theoretical sampling and involves deciding on which analytic grounds the sample should be chosen. Glaser and Strauss (1967) first advanced the idea of theoretical sampling as they discussed how to move from substantive to formal theory. Glaser and Strauss (1967) outlined four major strategies for theoretical sampling: maximizing or minimizing the differences either between groups, or between concepts in the data.

They provide this useful table to show the effect of the different strategies (shown slightly adapted, below, Table 2).

<table>
<thead>
<tr>
<th>Group Differences</th>
<th>Similar</th>
<th>Concepts in the Data</th>
<th>Diverse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimized</td>
<td>Maximum similarity in data leads to: Verifying usefulness of category; Generating basic properties; Establishing a set of conditions for a degree of category. These can be used for prediction.</td>
<td>Identifying/developing fundamental differences under which category and hypothesis vary</td>
<td></td>
</tr>
<tr>
<td>Maximized</td>
<td>Identifying/developing fundamental uniformities of greatest scope</td>
<td>Maximum diversity in data quickly forces: dense developing of properties of categories; integrating of categories and properties; delimiting scope of theory</td>
<td></td>
</tr>
</tbody>
</table>
The important thing to note here is that differences can occur at the group level, and also at the conceptual level. So, when theoretically sampling a similar group, this enables the filling out of categories if the data is similar, and helps with understanding of category variation and possible hypotheses, if the data is not. If theoretically sampling dissimilar groups, where the data is similar, this enables the theory to be built on key uniformities; if however, the data is very diverse, this helps to establish the limits of the theory.

While Eisenhardt (1989) discussed theoretical sampling for group differences, she did not consider theoretical sampling of concepts in the same way. She however talked about within case comparison based on categories, which can be seen as a variation of sampling for concepts in the data. The differences in strategy can perhaps be seen as a consequence of slightly different methodologies for theory building – Eisenhardt (1989) used cases while proponents of the grounded theory methodology rely upon “slices of data.” Regardless of these differences, we contend that theoretical sampling, used in a systematic fashion, would be of great assistance to the social media researcher trying to build theory over what can be a bewildering array of data sources.

While it is generally accepted that theories comprise constructs and relationships, there is not much discussion in the literature about the nature of those relationships and how they might be theorized. Relationships between constructs are not exclusively causal – there is a range of relationships that might occur when using qualitative data for theory building (Glaser 2005; Urquhart 2012).

Of course, discussions about theory building and the nature of theory did not stop with Eisenhardt (1989) or Eisenhart and Graebner (2007). Recent discussions in the Academy of Management Review, for instance, have revisited what might constitute a theoretical contribution (Corley and Gioia 2011), and the importance of combining existing theoretical lenses for building new theory (Okhuysen and Bonardi 2011). In the international business discipline, there has been a shift toward interpretivist theory building rather than toward theory refining case studies (Piekkari et al. 2009). In the social sciences, a typology of case studies has been proposed that encompasses the following; Subject (key, local or outlier), Purpose (intrinsic, instrumental, evaluative, exploratory), Approach (theory building, theory testing, illustrative/descriptive), and Process (Single: Retrospective, Snapshot, Diachronic, and Multiple: Nested, Parallel, Sequential) (Thomas 2011). We therefore conclude this section by noting that the case study method continues to be a vibrant and varied method, both in our discipline and others, and that theory building using case studies remains an important aspect of this method. In the next section, we examine an important building block of case studies of social media-related phenomena – the notion of a ‘digital text’.

**Defining a Digital Text for Social Media Environments**

One of the major issues confronting the social media researcher is the question of what might be a legitimate unit of analysis when a whole range of data in the environment might (and perhaps should) be studied. The social media researcher has to deal, for instance, with web pages, chat threads, emails, and many visual images.

What do we mean when we talk about ‘digital texts’? Schwandt (1997) gives an interesting overview of the different ways to see texts in qualitative research. He argues that Geertz (1983) proposed that to see social institutions, customs and changes as a somehow readable text is a potentially very powerful analogy. He then goes on to suggest that a narrower view of text is the hermeneutic one – originally biblical, classical and legal – but then that definition was extended by Ricoeur (1981) who declared that social action displayed some of the same features of a written text, and that the methodology of hermeneutics could then be applied to understand social action. Finally Schwandt (1997) points out that Denzin (1992) suggests, following Derrida, that everything – life experiences, events, relationships, activities, cultural artifacts etc. is a text. One issue here is that this idea seems to imply the primacy of language – what about images?

One of the key changes that has occurred in our use of the Internet in past years is the proliferation of images. For instance, thanks to the digital camera, we can all become producers of images that are then loaded instantly to web sites. While there is a long history of using visual data in qualitative research (Banks 2001; 2007), images tend to be analyzed separately, as opposed to being seen as part of a text that contains an image. So we consider that a working definition of a digital text would almost certainly
include images, with or without accompanying text. In the table below, we highlight some key characteristics of a digital text.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Held in digital format</td>
<td>Emails, chat threads, photographs</td>
</tr>
<tr>
<td>Contained on a web site</td>
<td>Web content</td>
</tr>
<tr>
<td>Co produced by more than one person</td>
<td>Web forums, wikis</td>
</tr>
<tr>
<td>Ephemeral</td>
<td>Comments on a link, a news feed in Facebook</td>
</tr>
<tr>
<td>Embeds other discourses</td>
<td>Link within a webpage, linking digital text to one another</td>
</tr>
<tr>
<td>Contains images</td>
<td>Avatar, web content, photographs</td>
</tr>
<tr>
<td>Contains video</td>
<td>YouTube clips</td>
</tr>
<tr>
<td>Lack of context</td>
<td>Microblog posts (e.g. “tweets” of 140 characters or less)</td>
</tr>
<tr>
<td>Linguistic innovations, blurring the</td>
<td>Emoticons, acronyms (e.g. “lol”)</td>
</tr>
<tr>
<td>distinction between icons and discourse</td>
<td></td>
</tr>
</tbody>
</table>

The first characteristic is perhaps obvious – a digital text, by definition, is held in a digital format. The digital format can be held in any number of file formats, which can sometimes cause issues around data management.

The second characteristic, we think, is a common one – many social media researchers will be studying a particular web site, or type of web sites, e.g. social networking sites, dating sites, gambling sites and so on. The question then is what sort of context does the web site or sites provide? Can we see the web site as providing an overall frame for the study? Can we thus conceptualize the web site as providing the natural boundaries for the case study? Or does the boundary lie beyond the web site?

The third characteristic – co-produced by more than one person – produces various ethical issues for the researcher. For instance, is the discussion on a public forum deemed as being in the public domain and therefore, like, say, a text from a newspaper? If the forum is membership only, and discusses sensitive issues, should the people on that forum know that they are part of a research project? What are our responsibilities when we too, participate in these forums? Gaining permission to use, for instance, a stream of collaborative ‘chats’ about a project within an organization is a different matter from considering what ethical concerns might operate in a web forum.

The fourth characteristic, the ephemeral nature of the text, also creates problems for the social media researcher. There needs to be a systematic way of capturing the texts so they are not lost. This can lead to the collection of large amounts of data, which in turn gives rise to two other problems – data management, and critically, deciding which texts might be worthy of analysis.

The fifth characteristic is that digital texts usually embed other texts through hyperlinks. A digital text is indeed usually not a “stand alone” text. One of the key features of digital environment is the ability to link content to each other. Therefore, digital texts are embedded in one another, creating new conditions for data collection.

The sixth characteristic, contains images, is, we think, an important characteristic. In such overwhelming visual environments such as those often afforded by social media (e.g. Facebook profile or pages) not to consider analyzing visual images that we encounter in such an environment may not do justice to that environment and its dynamics.

The seventh characteristic, contains video, underlines the previous characteristic. For instance, we know that video interviews give researchers much information from non-verbal cues. So we would suggest that
the social media researcher might also gain insights from such video sources and should actively consider where, when and how they can be incorporated into the case study design.

The seventh characteristic, lack of context, is also one we think that social media researchers should pay particular attention too. Regardless of whether one is analyzing a digital text, or crunching some ‘big data’, it is impossible to infer patterns without context. For instance, if a stream of Skype chats gets recorded, it might be also good for the researcher to note some basic context (e.g. time of day, day of week). Deeper context can be grasp by examining who are the people involved in these chats and, perhaps, by interviewing them.

Finally, the last characteristic is that of linguistic innovations and includes the growing reliance upon acronyms (e.g. “lol” for “laugh out loud,” “FYI” for “for your information,” or “IMHO” for “In my Humble Opinion”) in digital texts, as well as the tendency for digital “texts” to blur the distinction between iconic representations and discourse (Wolf 2000). Increasingly in web-based forums as well as in microblog posts, for instance, emoticons have complemented traditional language-based content and contributed to new ways of expressing complex ideas and sentiments online (Bos et al. 2007). They have transformed written text, making it in some ways closer to oral language, such as when an “lol” punctuates a digital text in a similar way that laughter punctuates an unmediated conversation (Spencer and Mandell 2007). Another similar linguistic innovation is the evolving use of hashtags on Twitter. From being simple tags on which a search can be performed, they also perform the function of a summarized and often humorous take on the tweet in question (Vaast et al., 2012).

Taken together, these characteristics of social media environments provide unexpected situations for researchers, who then need to adjust their theory building methods. Below, elaborate upon the challenges of case study design, data analysis and engaging in mixed methods.

**Challenge 1 – Case Study Design**

A key challenge – and the biggest change associated with the advent of research in social media environments – is the possibility to collect large datasets, much larger than what IS researchers could have hoped for just a few years ago. In this section, we consider how we might design a case study for a social media environment, given these large datasets. It is important to note here that we are defining a case study in broad terms, as we believe that a flexible definition of what a case is makes sense in social media studies. “In the sociological and anthropological literature, a case is typically regarded as a specific and bounded (in time and place) instance of a phenomenon selected for study. The phenomenon of interest may be a person, process, event, group, organization, and so on.” (Schwandt 1997, p. 12). This contrasts with the perhaps more customary definition used in the IS discipline that tends to see case studies as analogous with organizations (Myers 2008).

In particular, we consider how we might select data of interest within such large datasets, how we might decide on a case study boundary, how we might handle the issue of context, and whether single or multiple case studies are appropriate.

The availability of “big data” has arisen from two main conditions. The first one, obviously, and aforementioned, is related to the growth of social media, in particular in terms of available platforms and features as well as in terms of their increasing popularity that have contributed to new ways of communicating, working, etc. The second condition comes from increased computational abilities that have enabled companies to develop and populate huge databases as well as computer-mediated communications to become socially omnipresent and very cost-effective (Jacobs 2009).

From the possibility to collect huge data for research purposes, new challenges have arisen. In particular, how does the researcher select what might be significant data in a welter of logs, chats, emails and other data? The temptation for the unwary researcher is to collect as much data as possible – and of course, it is possible to collect a great deal of data in a social media environment. For instance, if collecting data on a virtual open source project, possible archival sources would include, for instance: emails; bug tracking data or software version control logs.
For illustrative purposes, let us consider a situation in which researchers decide that they actually do need to collect large volumes of data, say from social networking websites, blogs, or microblog archives. They are likely to rely upon an Application Programming Interface (API) that will help them access these data. One such publicly available API is Topsy, an openly accessible and relatively user-friendly search engine for Twitter, currently the most popular microblogging platform. Through Topsy, anyone, including researchers, can use the API to access the entire archive of Twitter (Topsy labs 2010). There are huge opportunities, of course, for researchers in being able to collect so much data and to be selective in their data collection (e.g. researchers can search for key terms and specific times). One problem here is that search engines and APIs can actually influence the data collection process: relying upon search engines and APIs can lead to biased samples, with the “bias” not always or easily understood by the researchers (Bollier 2010; Boyd and Crawford 2011). In effect, rather than purposeful theoretical sampling, the data collection can become the result of an uncontrolled and little understood process of opaque algorithm-led, data extraction. Therefore, just as a “sociomaterial perspective” (Orlikowski 2007; Scott and Orlikowski 2010) has recently led us to become more aware of the intricacies of the material and the social in the contexts we as IS researchers investigate, we also need to be aware of how these algorithms might influence our dataset.

So how might we begin to select data for our theory building case study? First, we might consider the notion of ‘central’ texts, and less central texts, and how those texts should be analyzed. There is always a ‘depth versus breadth’ concern in any qualitative research design; the more deeply a text is analyzed, the less texts will be analyzed. One possible route here is to analyze some texts more deeply than others – hence the notion of having central texts to analyze. Less central texts can be used to provide corroboration or triangulation. Readers may say at this point that this is merely primary and secondary data. We would beg to differ, because in fact, if theoretical sampling is used, ‘data slices’ could, and should, come from different sources – so we prefer the distinction between ‘central,’ and ‘less central.’ The logic of settling on a central text helps to provide an entryway for the researcher and to think about which data sources are important in the design.

We think that the idea of a digital text also has some reach when considering the unit of analysis of a study. Of course, the unit of analysis is always dependent upon the specific research questions of various projects. The idea of a digital text data forces researchers to think about what is the unit of analysis of their research, and what this unit of analysis means to them. For instance, in a recent study by one of the authors on bankers and online forums (Vaast, 2011), the unit of analysis was the post, but other units could have been selected (e.g. threads as a whole).

Once central texts, and less central texts, have been identified, we should then arrive at the boundary of the case study. One important consideration for social media researchers is whether their boundary is the social media boundary – do they study purely virtual environments? While it is entirely appropriate for ethnographers to study ‘virtual worlds’ as ethnographic environments, we contend that to place the boundary within or on a social media environment is not always advisable. First, when we consider the permeability of the division between ‘real’ and ‘virtual’ worlds, it is perhaps unrealistic to simply study what goes on in a social media environment. For instance, friendships in Facebook have corresponding relationships in ‘real’ life. There is often a reflexive relationship between the two. Second, one major challenge of studying social media environments, we contend, is that of context. How can we interpret a digital text if we are not aware of its context?

If we agree that context is important in a social media environment, whether when using digital texts or when using ‘big data’ expressed as visualizations, how do we build a context for our case study design? If we consider what our central text might be, we can at the same time consider how context can be built in. For instance, if we are interested in how hotel operators respond to negative reviewers on Trip Advisor, we might opt to use their responses on the web site as our central text for the case study (Scott and Orlikowski 2010). It might also be worth interviewing those owners about their response to follow up on issues that come up from the analysis or simply to find out about other factors that were in operation at the time for that owner. Similarly, if we collect a stream of posts about the development of a software product, it would be sensible to relate this to field notes about what was happening in the project at the time (O’Mahony and Ferraro 2007). Even if the researcher focuses exclusively on the virtual environment, the same argument for context still applies. For instance, let us imagine that researchers are investigating microblogging (e.g. activity on Twitter) in various contexts and situations (e.g. Heverin 2011; Jansen et al.
Researchers in this situation are likely to select single microblogging posts (i.e. the “tweets”) as their unit of analysis. At the same time, though, a tweet is very short (140 characters or less): making sense of a tweet in and of itself is difficult – tweets are to be understood within an ensemble (cf. the notion of “ambient awareness” (Boyd et al. 2010; Marwick and Boyd 2010). For researchers, the implication is that it is sometimes difficult, but important, to access the context of digital text data. Following our previous example, when studying microblogging, researchers could consider what was trending in Twitter at the time; they could consider the surrounding tweets, and the sequence of tweets and retweets (Boyd et al. 2010; Lotan et al. 2011; Namaan et al. 2011). In Vaast’s (2011) research on bankers and the online forums, the unit of analysis was the post, but posts were constantly resituated within their context of occurrence, i.e. the threads themselves. This problem of context is hardly unknown in qualitative research. For instance, researchers often need to situate an interview quote within the broader context of the interview. However, we note that digital text data might be more easily “lifted” from their context so it is especially important for researchers to remain aware of this issue.

While thus far in this paper, the plethora of data available to the social media researcher can be seen as carrying some disadvantages, there is one respect in which it can be seen as a distinct advantage; many digital texts lend themselves to theoretical sampling for theory building. The beauty of this solution is clear, as the developing theory determines which data ‘slice’ is examined next. In a social media environment, the ability to sample ‘data slices’ is extremely flexible. It will probably not require consent (although terms of usage for social media applications change and need to be checked regularly), or another phase of a study for ‘member checking’ as in a traditional case study.

As previously discussed, theoretical sampling can proceed in two ways. The first is exemplified in Eisenhardt (1989) where successive case studies are chosen on the basis of similarity or difference with the previous case and within case patterns. The second is a more systematic view from Glaser and Strauss (1967) where group differences are not only minimized or maximized, but sampling also unfolds with concept development. For instance, if the analysis of digital texts reveals many instances of a concept (e.g. self-presentation through Facebook status messages), one could choose to go on and sample more of that concept by interviewing individuals. Similarly, one could choose to sample further on a concept that has only occurred in a particular group of people, or is unusual (e.g. use of a Facebook status to communicate with only a few people in a very personal way). Glaser and Strauss suggest that sampling along diverse concepts quickly develops the theory and delimits the scope of the theory. Theoretical sampling in this systematic manner, we suggest, is underutilized in most management disciplines, including our own. Sampling using not only ‘different’ and ‘similar’ cases, but also looking for guidance in terms of concept development, means that we develop a much better, more grounded theory, with better scope.

Again, such flexible theoretical sampling of digital texts could be seen as a double-edged sword. When digital texts contain links to other digital texts that have some conceptual relevance, where should the researcher set the boundary for their case study? An obvious pitfall of is that researchers might soon find themselves overwhelmed by potentially relevant data. At the very extreme, link by link, the whole Internet could become their research setting, making theory building all but impossible and meaningless. There is thus an increased need to set boundaries for the cases under investigation.

We suggest however that if we consider social media environments to be an extension of social worlds that we already inhabit, then this problem is not what we think it might be. As we now accept the permeability between ‘real’ and ‘social’ worlds, we can see the social media environment as an extension (and sometimes a home for) various groups or organizations. So we would contend that, as long as theoretical sampling is carried out with a particular group in mind, the boundary of a case study should be easy to discern. For instance, a study on Facebook could potentially confine itself to a particular age group in a particular country. If a theory was then emerging about how that age group engages with Facebook, then some purposeful theoretical sampling of a similar age group in a different country, or one that has different membership of particular groups on Facebook, would: a) increase the scope of the theory (dissimilar groups), or, b) increase the density of concepts in the theory and establishing fundamental differences in explanation (similar groups but different categories). Additionally, if we chose to also to interview some people of a particular age or from a particular group, because of the need to further densify a concept, this would take the boundary outside the virtual one and into the ‘real’ world. As noted, in our view these boundaries are very permeable and should be dictated by theoretical sampling.

Challenge 2 – Data Analysis
Generally, when analyzing textual data, there are two main options: either the researcher can code the text at a detailed level, or they can apply a thematic framework of some kind. This of course still applies in the social media environment, but there are many more texts that can potentially be analyzed. So, distinguishing between these two options is important because they have different analytical ‘loads’; if analyzing a text in detail, there will be less chance to analyze many texts. This of course does not mean that, in practice, the researcher should not employ multiple methods of analysis – but that the time needed to apply different methods should be considered. In short, there may be a ‘depth versus breadth’ issue. As suggested in the case study section, we suggest that the social media researcher opts for some ‘central’ texts that can be analyzed in depth, because in the opinion of the authors, there is no substitute for such in depth analysis (with one important proviso – that this analysis is subsequently theoretically integrated).

When coding at a detailed level, there are three options for coding. First, bottom up coding, where codes are suggested by the data. Grounded theory method is a very good example of this approach, and can yield rich results because of a close tie with the data. Grounded theory method also gives a systematic way to theoretically sample different digital texts, as discussed previously.

Second, codes can be suggested by the literature and constructs used in instruments – this is ‘top down’ coding. Thirdly, codes can be applied which are a mixture of codes suggested by the data, and the literature. Overall, the coding approach gives the researcher assurance that what they have in their data really is in their data, because coding confers a more systematic approach to analysis. It also helps qualitative researchers defend themselves from charges of being selective about what they analyze.

Thematic frameworks are also frequently used to analyze qualitative data, and there are many possibilities available. For instance, there are many frameworks from discourse analysis and critical discourse analysis (e.g. Fairclough 1992 ) that could be applied to digital texts. Thematic frameworks or models built from relevant literatures, for instance a ‘sensitizing framework’ as suggested by Klein and Myers (1999) is also a frequent strategy used by researchers. An alternative is to build a thematic framework from the data, as suggested by Braun and Clarke (2006).

We would also urge social media researchers to seriously consider how they might analyze visual digital texts they encounter, and the role that such analyses can extend and enrich the emerging theory. There are various ways of analyzing such visuals. First, the text can be coded as if it was any other type of digital text. Given that most qualitative data analysis packages do have the capability to both store such visuals, and code them, there seems to be no obstacle to this type of analysis except a lack of familiarity, and a possible fear that this type of analysis might not be published. In fact there is already an example of such analysis, using grounded theory coding procedures, within information systems (Díaz Andrade and Arthanari 2009 ). We could also view images as not neutral, but as constructed texts (Banks 2007). This makes sense particularly with social media, where people manage how they appear on-line, and there are web sites, for instance, devoted to ‘photobombs’ where the subject of the photograph is sometimes accidently, sometimes deliberately upstaged by another element in the photograph. We can take a Foucauldian view of the image— who is doing the looking, whom does society empower to look at and be looked at, and what knowledge does this produce? Of course, visual analysis is not new, it has a long pedigree in cultural studies, for instance. One useful perspective comes from Ball and Smith (1992) who point out that it is important to distinguish between manifest and latent content for analytical purposes. For instance, an old photograph may contain a man with a mutton chop beard (manifest content) but the latent content concerns the social meaning of that beard at the time the image was produced (Ball and Smith 1992) in (Banks 2007). This, of course, brings us back to a key element which needs to be considered by the social media researcher when analyzing – context. This is why we would recommend that the issue of context be considered early, in the case study design.

Challenge 3 – Engaging in Mixed Methods

Social media environments also constitute unprecedented research settings, because they generate a wealth of digital text data that, technically at least, researchers may collect exhaustively and with little to no interference on the socio-technical dynamics at play. This new situation has generated tremendous opportunities (often epitomized in the enthusiasm for “big data”) but also challenges for researchers (Bollier 2010; Boyd and Crawford 2011; Lazowska 2008; Manovich 2011). Some of these challenges are legal, as detailed in Allen et al.’s (2006) seminal essay on regulatory implications of automatic electronic
data collection. Other key challenges that researchers have been aware of are of an ethical nature, and deal
with, for instance, the blurring of the distinction between what is private and what is public, or questions
regarding whether electronic data are “people” or “public” (Buchanan 2010; Buchanan and Ess 2008;
Hudson and Bruckman 2004; Schultze and Mason 2011; Sveningsson 2004). Aware of these critical
issues, in this essay, we focus on another important, and so far less discussed, challenging implication of
“big data” for theory building, the need to engage in mixed methods.

In this regard, Anderson (2008) provocatively called for a drastic change in the scientific method, given
the rise of electronic data becoming publicly available to many. He predicted that “the end of theory”
was being brought about by the “data deluge” of the Internet, search engines, and social media. Intrigued by
this provocative thesis, we however believe that the scientific community can still contribute to society not
only by discovering and explaining correlations, but also by theorizing causations, and multiple-order
effects between concepts. Others have discussed how electronic data might transform both the process of
developing new theories and the resulting theories. John Seely Brown, for instance, as reported in Bollier
(2010, p. 8), addressed how newly available data can help researchers discover “generators” for new
theories if researchers are able to make sense of what in their data, corresponds to “outliers” and what
reveals meaningful patterns:

“How can you invent the ‘theory behind the noise’ in order to de-convolve it in order to find the
pattern that you weren’t supposed to find? The more data there is, the better my chances of finding
the ‘generators’ for a new theory.”

In this paper, we consider that, for researchers to be able to discover the “generators” for a new theory,
they need more than powerful computational abilities, although they obviously do need such resources to
be able to sift through huge volumes of data. To do so, researchers interested in theory building, and not
just in descriptive analyses or in testing existing theories with electronic data, would do well to engage in
mixed methods analyses.

The huge volume of data that researchers can collect in social media environments creates data
processing, reduction, and management challenges. Obviously, such a volume of data can be impressive,
but is also utterly meaningless in and of itself. Collecting huge volumes of data, per se, does not guarantee
a theoretical contribution for the research. If anything, big data might make building theory more
difficult, because theorizing patterns and outliers becomes especially challenging in a wealth of
decontextualized data. Researchers might not always feel well equipped to handle these datasets.
Researchers who are usually engaged in qualitative research might find themselves intimidated by the
volume of data, and might not know how to “attack” it to develop, for instance, a grounded theory of their
phenomenon of focal interest. Getting deeply into a corpus of semi-directed interviews of reasonable
duration is already formidable enough; how to make sense of years of archives from online discussion
forums, for instance? For qualitative researchers dealing with such new contexts and new data, then,
being able to delve into data from a more quantitative angle, thanks to descriptive visualization
techniques, can be helpful.

Quantitative researchers are not immune to the challenges of big data for theory building either. A major
issue for quantitative researchers is that, when huge volumes of data have been collected, the significance
level loses some of its meaning, and putting too much weight on it for theory justification purposes, as
opposed to the power of the analyses, might weaken the resulting theory (Bollier 2010; Boyd and
Crawford 2011). To build theories of these new environments, quantitative researchers therefore cannot
rely singlehandedly on well-established statistical indicators, and need to be able to give meaning to their
data. Adoption of more qualitative approaches to theory building may assist with achieving this
sensemaking.

Therefore, handling digital text data for theory-building purposes might force researchers out of their
typical (qualitative/quantitative) comfort zone, and lead them to embrace methodological creativity.
Qualitative researchers might try to deal with larger data sets with descriptive quantitative methods, and
quantitative researchers might delve into the meaning of their data, rather than focus mostly on the
significance level of their statistical procedures. Obviously, calls for triangulation for better theory
building are far from new (see, in organization research, (Jick 1979; Shah and Corley 2006; Van Maanen
1979), as well as in IS (Kaplan and Duchon 1988; Mingers 2001; Mingers 2003)). We believe that building
theories with case studies of social media environments especially requires that IS researchers adopt such
mixed methods. Triangulation, in the sense of combination of different methods, especially involving qualitative and quantitative aspects, is essential for theory building case studies of social media contexts. For one, it can help researchers deal with huge volume of data and sift through them, relying upon multiple techniques, to identify patterns to be theorized. Moreover, mixed methods offer researchers the ability to develop complementary perspectives on the same phenomenon and to discern previously hidden connections among concepts.

On a practical note, many of us, in the IS research community, are more familiar with qualitative or quantitative research, although many of us would also not consider ourselves exclusive proponents of a single method over other ones. Engaging in multi-method theory building research projects can help us reach beyond what we have usually become more familiar and, for some of us, even expert on, methodologically. This, obviously, makes such research demanding, because it is not enough to develop “qualitative” or “quantitative” research, and researchers have to develop research projects that combine various methodological “ideal types.” We therefore hope that the opportunities and challenges of developing mixed methods for theory-building purposes will lead IS researchers to reach out and develop collaborations across the traditional methodological divides. Researchers at ease with the latest computational methods can summarize vast amount of data and develop impressive, descriptive visualizations (Lazowska 2008). Other researchers, more attuned to discovering theories, could work with them in order to make sense of empirical patterns and soundly the “creative leap” (Langley 1999) that is necessarily involved in any theory building effort. Such mixed methods research projects would still be challenging to develop, because there are still currently fewer established criteria to evaluate mixed methods than for more traditional methods. We would therefore like to see researchers investigating social media environments who are involved in mixed methods projects be more explicit about their methodological choices, and engage in conversations about those choices. This would be most productive for the IS research community, as argued in Majchrzak (2009) and Te’eni (2009).

Conclusions

The purpose of this paper was to reflect on emerging experiences of building theory of social media-related environments, and to articulate the challenges and opportunities that IS researchers are likely to encounter in these new environments.

Theory building remains an enduring challenge for the IS discipline, and more generally for all disciplines that deal with ‘applied’ environments. A more systematic and conscious use of theoretical sampling, along with developing concepts, would be very beneficial in this regard. This means going beyond the ideas advanced in Eisenhardt (1989) about theoretically sampling similar and dissimilar cases, to thinking about sampling cases on the basis of concept development. This enables the theory to either be further instantiated in various cases, or broadened out in scope, but along particular constructs.

The advent of social media has ushered in two major changes in how we use the Internet. First, it is a much more collaborative era, that allows people to state their opinions, and for their opinions to be gathered, via crowdsourcing. This gives the researcher an increased opportunity to study varying viewpoints, and gain corroboration of particular viewpoints. Second, it is a much more visual era; people use digital cameras, upload images, but also have control over those images in an unprecedented way. Academics in general have been slow to respond to these new ‘visual’ texts, other than colleagues in cultural studies or art history. Visual images are now so much of our social worlds in social media that we think we could, and should, as researchers, be analyzing those images.

The advent of ‘big data’ cannot be understated as a development either; there is a huge potential to collect a large amount of data, but with a corresponding need to make sense of that data. The role of context in a social media should in this regard be acknowledged, as it is required to make sense of either a large data set or a ‘slice’ of data. We recommend then, that, when embarking on research in social media environments, there be an active consideration of how context will be accounted for.

Finally, we considered that ‘big data’ and the need to analyze and contextualize these datasets has generated a critical need for researchers to go beyond their familiar and comfortable methodological choices, and to embrace mixed methods. Engaging in mixed methods is likely to trigger productive collaborations among researchers and to help generate thought-provoking conceptualizations.
This paper represents a first step in opening a dialogue among IS researchers interested in theorizing about social media-related environments. We look forward to debating with our colleagues, and to further elaborating on these ideas about possible methodological approaches. Important questions tightly related to these issues are still open, and are worth considering in the community as a whole. One such crucial question for many of us is that of the publishability of highly innovative research on social media in high-level academic journals.

We conclude by emphasizing the important role that the IS discipline can play in researching and theorizing social media. These new research directions hold promise for the IS discipline to remain or become more relevant to practitioners (Straub and Ang 2011), as well as “reach out” to and develop its influence on other social sciences such as sociology or political sciences (e.g. Wattal et al. 2010). IS researchers have a deep, and arguably unique, understanding of technical and social / human dimensions of dynamics associated with social media. Moreover, IS scholars have become attuned to the challenges of investigating Information and Communication Technologies and to the need to theorize them. They are thus extremely well placed to inspire other disciplines to tackle the investigation and theorizing of social media-related environments and dynamics.

References


Research methods


Research methods


Research methods