ZAGADNIENIA
NAUKOZNAWSTWA
4 (210), 2016
PL ISSN 0044 – 1619

Maria Nawojczyk[*]
Jarosław Królewski
Faculty of Humanities
AGH University of Science and Technology, Cracow

Using *Big Data* in Innovation Research

Abstract. One of the major contemporary trends revolutionizing social-sciences computing is, *inter alia*, the so-called *Big Data* effect, meaning fast and multidimensional analyses of large volumes of data. Technologies related to *Big Data (Volume, Velocity, Variety)* have considerable impact on the tools of contemporary sociologists, providing them with immense data resources in real time. *Big Data* is a term encompassing all data, the analysis of which may bring quantifiable benefits, not only in terms of business but also in science and research. Modern technologies change and greatly impact the methodology of research conducted, giving rise to numerous questions and doubts both strictly methodological and ethical in nature. One of the main challenges related to *Big Data* is the possibility of using large data agglomerates as early as at the stage of conceptualizing and operationalizing the subject of social-sciences computing. The possibility of transforming *raw* data into pieces of information, and then into knowledge, may soon become an indispensable and desirable element applied in social engineering in establishing the practical applications of research and in predicting future social phenomena. The latter could be particularly useful in such an important and sensitive field as innovation research. Two cases of innovation: a social and technological ones are discussed in the paper. Using these two cases we will present a model of analyzing innovations in real time. The proposed model is a new approach to study innovations.

Keywords: *Big Data*, social sciences computing, analysis in real time, social innovation, social change, model to study innovations, Mixed-Methods Research

Używanie *Big Data* w badaniach nad innowacjami

Abstrakt. Jednym z głównych, współczesnych trendów przynoszących rewolucyjne zmiany w informatyce społecznej jest efekt *Big Data*, czyli szybkiej wielowymiarowej analizy wielkich zbiorów danych. Technologie związane z *Big Data* mają znaczący wpływ na narzędzia badawcze jakimi mogą posługiwać się współcześni socjologowie. Dają im one możliwość dostępu do źródeł danych w czasie rzeczywistym. *Big Data* to termin odnoszący się do wszystkich danych, których analiza może przynieść policzalne efekty, zarówno w kategoriach ekonomicznych jak i badawczych. Nowe technologie wpływają w znaczącym stopniu na metodologię badań, przynosząc szereg pytań i wątpliwości tak natury metodologicznej jak i etycznej. Jednym z wyzwań związanych z *Big Data* jest możliwość posługiwania się wielkimi zbiorami danych na wczesnych etapach konceptualizacji i operacjonalizacji problemów i hipotez badawczych. Przekształcanie surowych danych w informacje i wiedzę stanie się nieodłącznym elementem nie tylko inżynierii społecznej ale również praktyki badawczej dającej lepsze możliwości predykcyjne naukom społecznym niż to miało miejsce dotychczas. Te możliwości wydają się szczególnie użyteczne w badaniach nad innowacjami. Przedstawiamy je w szczegółach na dwóch przykładach innowacji – jednej technologicznej i drugiej społecznej. Przykłady te służą nam do zaprezentowania modelu, który stanowi nowe podejście do badania innowacji

Słowa kluczowe: *Big Data*, informatyka społeczna, analiza w czasie rzeczywistym, innowacje społeczne, zmiana społeczna, model badania innowacji, łączone metody badawcze

[*] Address for correspondence: Faculty of Humanities, AGH University of Science and Technology, ul. Gramatyka 8a, 30-071 Krakow, e-mail: maria@list.home.pl

1. Introduction

The discussion and analysis of a new type of society has a well-grounded history in last few decades in sociology. Considering three concepts of the emerging society, as a post-industrial society (Bell 1999; Tourain 1974), an information society or network society (Castells 2000), or one specific to economy (a knowledge-based economy), all three of them stress the role of innovations as a driving force for development. The impact of particularly technological innovations on economic development was present, observed and described in previous stages of social history, mostly in the industrial era. Such preconditions as an institutional framework, social and cultural settings, the structures and strategies of business firms as well as the opportunities and constraints of the available technologies were categorised in different techno-economic paradigms connected with the waves of technological revolutions (Dosi, Orsenigo, and Labini 2005). But as Joseph Schumpeter wrote in his classic text, development is defined by carrying out new combinations which are appearing discontinuously (2000). In this classical approach innovations are individual phenomena happening from time to time in particular places and times, which are then absorbed and diffused in a broader scope, so first creative destruction occurs and then the system moves toward equilibrium (Schumpeter 2000).

The classic approach to innovations is no longer relevant in modern societies where innovations are present in all spheres of social life and the status of the economy depends directly on their presence. Innovations in well-developed countries become a necessity, subject to state policy, and an indicator of economic advancement. They still occur within the framework of a variety of techno-scientific innovativeness systems. These systems differ among themselves based on social, cultural and legal settings they occur in. In the same time these system consist a field of study how to apply innovative approach to the study of innovations.

In order to understand this type of systems, an analysis of their social infrastructure is needed. Examples of how many and how different indicators are taken into account to describe this social infrastructure can be found by analyzing the global innovative index (GII 2014). One efficient way to become more innovative is implementation of laboratory innovations in a number of institutions clustered around research centers. This leads to a spatial concentration of scientific, technical and business organizations with excellent examples found in the USA, e.g. Silicon Valley or Route 128 (Boston). Many countries have done this successfully, judging from the data of global competitiveness report where 37 out of 148 countries are categorized as innovation-driven economies (GCR 2014).

The trend to accumulate resources to enable dynamic development of techno-scientific innovations has also had its influence on urban projects. We can observe more and more frequently the transformation of whole urban agglomerations (both in the area of spatial development, transport, recreation, sport and culture, and the

architecture of specific buildings) to meet the needs of the creative class, including workers in the high technology sector (Florida 2001).

It is symptomatic that the spread of innovations has blurred the borders between engineering and social technologies (Beinhocker 2007) as well as between material and non-material commodities (Ritzer 2004). There are at least two important dimensions in the debate on the knowledge-based economy: the growing importance of information technology in socioeconomic life and the nature of information as commodity (Tonkiss 2006). We will focus only on the former in our analysis. We will propose a model based on four dimensions to study innovations in real time using *Big Data*. Providing two examples we will show advantages and disadvantages of the proposed approach. We will also argue to what extent this approach fits in the debate on innovations study so far, and how it opens new possibility in this field of research.

Innovations in contemporary societies are multi-level phenomena socially and culturally embedded. Therefore, following the notion of social entrepreneurship (Steyaert, Hjorth 2006), and pointing to connections between that entrepreneurship and innovations we will treat innovation as a complex social creative process that influences, multiplies, and transforms the context within it was grounded. That implies a mixed-method approach in which we can combine micro and mezzo levels of analysis accommodating individual actors and organizations in constitution of different fields (Bourdieu 2005). It will also allow us to compare different strategies of these actors in different institutional, spatial, and social settings (Fligstein, McAdam 2012). No matter how heterogeneous the process of innovation might be, the outcome should be in the form of social change. We will understand the social change as an institutionalization of new social practices. So, we will try to show how useful the new information technologies are in the measurement of social change.

2. Modern technologies and their impact on new types of data

Modern information technologies have made us look at social life today as something very different than it was a few decades ago. Human interactions have changed along with the formation of not only complex systems of social roles and the evolution of the broadly understood public communication, but also of such concepts as power, trust, risk, safety, and the ability to adapt to the new technological environments (West, Turalska, Grigolini, 2014). Looking for inspiration, enabling us to fully present the most current technologies associated with the processing of immense amounts of data, and then using them for strictly scientific purposes, we first looked at the solutions based on algorithms merging data from different sources. *Big Data*, understood as a phenomenon signifying primarily technological and social capabilities of processing and extraction of immense volumes

of data, is currently one of the most popular concepts emerging among those in the technology industry and analysts from around the world. Presented by many as an opportunity for the development of new theoretical models and empirical research, it is also heralded by the McKinsey Global Institute as the next wave of technological revolution.[1]

There is no doubt that the potential of *Big Data* is both a great challenge and a great unknown for both business and science. However, there is a difference between using big data in scientific analysis and in commercial applications. This difference lies in the goal of analysis, which is in academic research the generation of abstract knowledge, and in commercial applications is based on the manipulation of behaviour (Schroeder, Cowls 2014). Access to databases, public and, in many respects open, API (Application Programming Interface) of systems such as Facebook, Google and Twitter allows us to build research apparatus and tools that use powerful volumes of data. Web 3.0 provides experts and sociologists with data from the *multi-screen world* in real time. We can instantly monitor the content from online forums, analyse the sentiments of users' statements, and conduct ethnographic research based on millions of citations in social media obtained via *web crawlers* and specific queries. When we add to that the almost instant access to information on trends (Google Trends) and services related to combining different API systems (IFTT), we will come up with endless possibilities supported by the IT infrastructure thus far inaccessible to the social units.

Systems that use the *cookie* mechanism are able to very solidly and permanently collect data on specified services and products that are used by us, advise us in matters related to love, health and holidays, and affect our political preferences or civic activity. Mobile applications and a wide access to the Internet around the world allow us all to have the same world within our grasp. Today immense volumes of data are generated by ourselves. Social practices could be disrobed us network exhibitionism, and the protection of our personal data, which would seem crucial, has lost the battle with convenience. This is because for *Big Data* research the phenomena under investigation are digital platforms and peoples' digital traces.

New relationships and phenomena appearing on the Internet are the subjects of many scientific debates. There are questions about the knowledge that flows from the tools extracting immense amounts of data here and now. The potential of social and demographic data regarding the behaviour and psychology of the crowd which large organizations wield in their hands is almost infinite, and it allows the conducting of many social experiments, including those related to innovation and

[1] McKinsey Quarterly, *Ten IT-enabled business trends for the decade ahead*: http://www.mckinsey.com/insights/high_tech_telecoms_internet/ten_it-enabled_business_trends_for_the_decade_ahead [access: 22.08.2014]. However the business intelligence techniques existed before, we observe now broader scope and strategic use of them.

social engineering. In these debates to some extent we can find traces in common with the debate over application of process-produced data in academic research.[2] From a methodological point of view with both types of data (big data and process-produced data) we are facing problems of production and selection biases. In both cases we have to deal with disassociation of data collection and data analysis as well as easy access to them (Baur 2009). However, there is one important difference between these two types of data: the process-produced data was and still is produced on purpose, while digital traces in *Big Data* are left unintentionally. Therefore, taking into account production bias we have to look into societal and institutional settings in the case of process-production data, but into technological and cultural embeddedness when dealing with *Big Data*. This technological and cultural context of gathering data should consider available technology and social practises of using technology in studied societies.

In many fields of research investigation we are combining elements of qualitative and quantitative methods of collecting data, and using triangulation we are trying to obtain better understanding of analysing problems. In many of these mixed-method approaches secondary data are included (Baur 2011). The above-mentioned availability of mass data and big data presents sociologists with the tough task of understanding the phenomenon as such, the hazards emerging from them, and the methodological awareness which should be applied in research conducted on the basis of them. Following that, we will try to demonstrate the advantages of using *Big Data* in innovation research.

3. How is *Big Data* employed in innovation research?

Focusing on an analysis of *Big Data* in the context of research, we will consider social innovation very broadly, including both economics and other topics related to the change of paradigms and systemic approaches with regard to changing social reality ("reshaping society"). Innovations on the basis of sociological analysis are often considered as essential components of social progress. Social innovation does not necessarily need to be linked to the achievement of economic success and may simply be a new solution to problems known to us already. Many projects and products are intended not only to achieve financial success, but also to change people's habits and attitudes and consequently lead to the formation of a specific type of society that is increasingly influenced by strictly designed innovation. There is no doubt that the capture of a social change caused by planned actions or spontaneous activities is one of the most reliable indicators of verifying the existence of innovation. In our case studies we will refer to both starting points

[2] See the special issue of HSR, Social Bookkeeping Data, 2009 (3).

of innovation (planned actions and spontaneous activities) to show that our model is able to capture a varied types of innovations.

In our definition of innovation, social change, understood multidimensionally as a change in behaviour (e.g. introduction of mobile devices causing reorganization of our existing habits), change of attitudes (e.g. causing personality changes), and a change of roles and systems of social interactions (e.g. redefinition of social relations, changes in organizations, or on the ladder of prestige) is critical. The second most important issue relevant to the process of research on innovation is the concept of diffusion of innovations, in the economic meaning of scaling innovation. Taking into account both crucial characteristics we can look at innovation in three dimensions (Schmitt 2014):

a. non-technical innovations in an organizational context,
b. social innovation to be connected to technological innovation,
c. social innovation as new social practices.

Thanks to modern IT innovations, the time of creation and validation of innovations has also shortened. Today, virtually anyone can quickly and cheaply validate their vision on a global scale without investing large financial resources. This is possible through the use of web and mobile technologies. Contemporary modern technologies (often being innovative themselves) allow in-depth and immediate verification of the effects of actions related to social engineering – design of social innovation (Schmitt 2014). The innovation lifecycle is very short. It is difficult to capture the point at which the innovation actually occurs. Today's modern technologies, using the Internet and systems to track in real-time social reactions to the innovative phenomena, allow us to explore innovation more effectively and adequately, both from a business and a sociological point of view. For the purposes of this study, we propose to identify and measure innovation in the model based on four dimensions:

1. **degree of conformance of initial assumptions and anticipated results:**
   a. *What is a problem as defined by an innovation leader/author?*
   b. *How do we want to resolve it?*
   c. *How is the problem finally resolved?*
2. **the degree of susceptibility of innovation to change initial assumptions**
   a. *What are the additional assumptions and secondary results after introducing innovation?*
   b. *How much does the public opinion influence innovation?*
   c. *Can innovation survive after a change of the initial model?*
3. **the degree of diffusion of innovation (virality):**
   a. *What do people think about the problem?*
   b. *What do people think about innovation?*
   c. *How quickly and in which direction is innovation spreading?*

4. **the degree of institutionalization**:
   a. *What is the form of institutionalization?*
   b. *How stable is institutionalization?*
   c. *What kind of fields (social, political, economic) does institutionalization affect?*

We will develop detailed analysis of their application on the real examples in the next section.

4. Four dimensions for measuring innovations – real-time research

The approach to investigating innovations in real-time research allowed us to overcome categorizations of innovations. Even if we do not believe that purely technological innovations exist, we can still find differentiation between technological-economic and social innovation in existing studies. We consider this division as no longer relevant but to strengthen our arguments we will use them in two case studies – the first technological and planned (*Sherly*), the second extremely social and spontaneous (*freedomapples*).

4.1. Degree of conformance of initial assumptions and anticipated results

One of the methodological problems with this dimension is connected with the time of observation of innovation "in the making". With many innovations we as observers and researchers are not able to pinpoint the beginning of the innovative idea in real-time research. It is obviously easier with innovations based on a business plan and not appearing spontaneously. It is much easier to discover that we are dealing with innovation at an advanced stage of the process. Therefore an exploration of innovations could be based on two possible perspectives of research, which could be carried out separately:

a. Identifying a problem, looking at and exploring many aspects connected with the analysed issue – one of them could lead to innovation. We will apply this approach to the case of *freedomapples*;
b. Identifying/capturing and observing developing innovation in order to judge its success. This will be applied to the case of *Sherly*;

Sherly

In recent years there has been rapid development of new Polish companies, especially in the areas related to the ICT industry. Young startups are supported mainly by seed funds and venture capital institutions. One of the ideas that has gained favor with investors is a project called *Sher.ly*. The main concept of the

Sher.ly team was to develop an alternative private software solution as competition to the public cloud computing method for storing and synchronizing files with storage devices. *Sher.ly*, an SaaS-based software, creates a local file cloud on on-site storage to share and collaborate on sensitive data. The origins of the company date back to the beginning of 2013.[3] The company was founded by two experienced specialists from technological (especially experienced in secure transfer technologies) and marketing industries. The company's main goal is to protect the confidentiality of sensitive data sharing with secure access control.

Among the anticipated results we can identify two main ones. One is to replace cloud solutions with private group file exchanging (turning public clouds into private) with communication between groups working on different sets of files, and the other is to succeed in the B2B market (software solutions for companies). We would like to stress that both results of this technological innovation have social consequences. Thanks to access to social media monitoring tools we had opportunity to check the activities, opinions and writings of the creators of *Sher.ly* available on the Internet.

Image 1. Błazej Marciniak, CEO of Sher.ly posts about security on Twitter

Freedomapples

Social media (facebook, twitter, linkedin) have become the most engaging pastime for Internet users. Those kinds of platforms have given us the opportunity to create a number of social, economic and political initiatives. Many of them bear the hallmarks of spontaneous innovation, of initiatives exhibiting characteristics of a social movement. An interesting example of social innovation, lately appearing on the Polish Internet and spreading throughout the country, was a spontaneously constructed action resulting as a protest against the imposition of an Russian embargo (June 2014) on Polish fruits and vegetables. A Polish journalist launched via twitter post (July 2014) an internet campaign to "Eat apples against Putin," which called for more frequent consumption of domestic apples. Poland is one of the leading producers of apples in the world. More than half of that production

[3] Sher.ly Inc: http://www.crunchbase.com/organization/sher-ly#sthash.dkFWAXLT.dpuf [access: 22.08.2014].

was exported to the Russian market.[4] One in three Poles bought apples during the campaign not only for their taste and nutritional value, but for modern economic and patriotic reasons, following the Internet campaign.

Image 2. Freedomapples meme examples

Source: *Join '#Jedzjablka' Campaign against Russian Embargo on Polish Apples:* http://inside-poland.com/t/join-jedzjablka-campaign-against-russian-embargo-on-polish-apples/ [access: 20.08.2014].

In both cases the problem they address was clearly defined – in case of *Sherly* by business plan, in case of *Freefomapples* by political and economic situation. Also the resolutions of problems were visible – one in again in business plan, the second appeared as spontaneous reaction of society. Both provided to final outcome but in first case unexpected and in second exceeded expectations. The analysis of this processes are presented below.

4.2. The degree of susceptibility of innovation to change

We assume that every innovation process occurs within a particular social climate and depends on social and cultural backgrounds, some of which are more favorable to innovation and some less. We are treating this characteristic as given to the environment in which the innovation happens. From observation of the dynamics of the innovation process we can thus indirectly judge features of social climate. The analysis of dynamics of the innovation process in real time gives us the opportunity to spot the inflection point, which is particularly helpful in the prediction of an innovation's outcome. Another issue is connected with secondary results. We can observe secondary results from the beginning, but we are not usually able to point to which one will lead to an innovation shift[5] and bring different outcomes.

[4] Newsweek. *Measurable success apples against Putin*: http://biznes.newsweek.pl/sukces-akcji-jedzjablka-przeciw-putinowi-newsweek-pl,artykuly,347030,1.html [access: 22.08.2014].

[5] PIVOT – *structured course correction designed to test a new fundamental hypothesis about the product, strategy, and engine of growth* (Ries 2011, 103).

Sherly

An additional assumption which appeared during work on Sher.ly was the leveraging of existing storage infrastructure. Customers would avoid paying for syncing so the Sherly team decided to launch a new data storage product, *SherlyBox*. The secondary results of *Sherlybox* was success both on the B2C market (crowdfunding) by introducing a new product, and a high rate of virality on the Internet.

The idea was the result of the need for security, as reinforced on Internet users by media reports of threats to their online experience. The origins of the company correspond to the scandals associated with Edward Snowden's disclosure of confidential information relating to the surveillance of Internet users by the National Security Agency in the US. Sher.ly is opposed to cloud technologies where many people store large amounts of data without knowing for sure who actually has access to the data. After preparation of a Minimum Viable Product and considering other factors (e.g. market validation, social media research, public opinion) the team decided to prepare an additional hardware product (SherlyBox) to complement the designed software. The idea for the new product was supported by worldwide debut of iBeacon technology, which is based on Bluetooth Smart Technology and enables energy-efficient communication between different devices. The Sherlybox pivot was proof of the high susceptibility for innovation to change under the influence of public opinion. The second generation of the product has been successfully crowdfunded. Its Kickstarter.com campaign showed that Sherlybox, the unplanned new product, created a significant competitive advantage which the software alone could not achieve.

Freedomapples

What was initially an Internet challenge activity (people posted pictures of themselves eating apples) and an individual political demonstration turned out to be, in less than a few weeks, an incredibly successful international action widely commented on in the world media. The most striking example was the initiative of the British magazine "The Economist," encouraging their countrymen to support the Polish attitude. The journalist created the hashtag #freedomapples and published a list of places in London to buy Polish apples. So the change in this case was connected with the public response leading Internet action toward organized movement.

4.3. The degree of diffusion of innovation (virality)

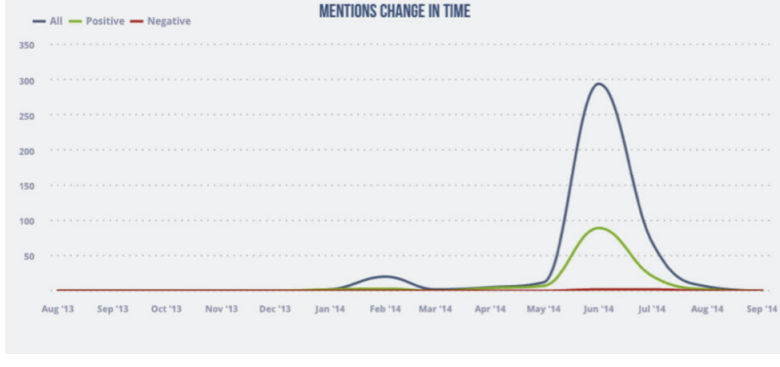
The degree of diffusion can give us considerable interesting information about both the spatial range and the social range of a given innovation process. The pattern of diffusion could be a subject of study in itself. This part of the innovation

process is much better described than any of the others. This is also the part which could be more precisely measured with the help of Internet tools. As a point of reference we would like to invoke the notion of *innovation diffusion* characterized by five indicators introduced by Everett M. Rogers (relative advantage – *the degree to which innovation is perceived as better that idea it supersedes*; compatibility – *the degree to which an innovation is as being consistent with the existing values, past experiences and needs of potential adopters*; complexity – *the degree to which an innovation is perceived as difficult to understand to use*; trialability – *the degree to which an innovation can be experimented with on a limited basis*; and observability – *the degree to which the results of an innovation are visible to others*) (Rogers 2003, 229–258).

Sherly

As regards Sherlybox, the diffusion of new product innovation was extremely high. On Kickstarter, the team reached $154,106 pledged (of the $69,000 goal) from 896 supporters. After the Kickstarter campaign the company achieved exceptional marketing success. The widespread interest of investors from around the world focusing on SherlyBox confirms the significant transformation of the product.

Image 3. Social media monitoring – SherlyBox debut



Source: *Query results*. Sentione.com [access: 20.08.2014].

Freedomapples

The campaign has become a symbol of economic patriotism resulting in innumerable references, content, articles, photos, and videos, all spontaneously generated by the Internet community. Following the Internet success came the spontaneous new social practices in real world. In many cafes and restaurants apples were added to the menu and were given to customers along with the bill. Many firms (banks, shops, offices) replaced their offered sweets with apples.

Image 4. Social media monitoring – Freedomapples

Source: Kozak Izabela. *The second life of Polish apple:* http://socialshake.pl/jedzjablka [access: 22.08.2014].

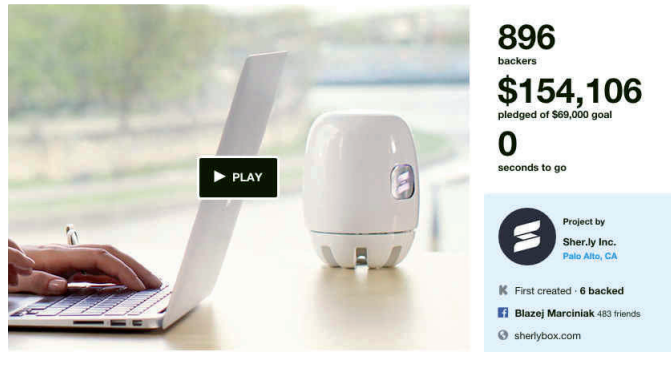
4.4. The degree of institutionalization

The outcome of the whole innovation process finds its confirmation in institutionalization practices, financial investments, accumulation of social capital, and changes in social practices. Sociology is providing us with tools of analysis and measurement for each one of them. In cases like *Sherly* it can be simple economic measurement followed by market research and analysis of the new product. In the case of *freedomapples* the outcome has a multidimensional character and should be analyzed accordingly using mixed methods of institutional analyses.

Sherly

Agreement between the Sher.ly team and its backers was institutionalized after a successful founding. Sher.ly is going to deliver products to the supporters. So in this case micro-investments legitimize the idea of Sherlybox in a highly measureable way.

Image 5. Sherlybox kickstarter.com success



Source: *Sherlybox – turn your public clouds into private & unlimited*: https://www.kickstarter.com/projects/sherly/sherlybox-a-private-and-shareable-cloud-on-your-de [access: 22.08.2014].

Freedomapples

Observing the evolution of the campaign, it is worth noting that the action went beyond the online community, was quickly picked up by the local Polish companies, institutions and international corporations (institutionalization), and led to changes in social practices. The process of institutionalization was and still is observable in many spheres and on different political, economic and social levels. The issues of apples and apples as such were used in political campaigns for local elections. State subsidies were introduced to producers of apples, as well as new regulations which make it easier and more profitable to supply goods to various charitable foundations. Programs for healthy food were introduced in many Polish schools, replacing sweets with apples. Many companies following earlier spontaneous replacement of snacks with apples have now brought it under regulation (see the example of Polish Railroad Company). Eating apples has become fashionable.

Image 6. Example of "freedomapples institutionalization" by the Polish national rail operator

Source: *"Eat apples" in the train. Railway joins the action. Travelers will distribute 40 tons of fruit*: http://pieniadze.gazeta.pl/pieniadz/1,136158,16590358,_Jedz_jablka__w_pociagu__PKP_dolacza_do_akcji__Rozda.html [access: 20.08.2014].

The table below summarizes our investigation into two cases supporting our arguments that in innovation research we do not need to place the cases into technological and social categories: they go through the same phases, and could be analyzed by the same tools.

Table 1. Four dimensions for measuring innovations – real time research

Project	*Sherly*	*Freedomapples*
Problem	IT security	Economic sanctions
The initial point	Business plan (formal); Kickstarter (semi-formal); Social media (informal)	Social media (Informal via tweet)
PIVOT	Sherlybox	Not defined
Virality	Media impact	Media impact
Institutionalization	Financial success on kickstarter.com – backers	Apples in companies and institutions

5. Pros and cons of using *Big Data* in research processes

The importance of *Big Data* in the modern world and research is undeniable and it will continue to grow. Today sociology is understanding *Big Data* as not only massive data sets collected automatically but also "'user-generated content', or information that has been intentionally uploaded to social media platforms by users … their tweets, status updates, blog post and comments, photographs and videos and so on" (Lupton 2015, 3). Both presented cases reflects analysis based on *Big Data*.

The general discussion covers different aspects of risk society, technological rationalization versus social, and technological determinism with a growing anxiety that individual behaviour can be not only predicted but manipulated by the new technologies. Therefore in public opinion methods of analysis and collection of large aggregates of data are seen mainly as a threat to autonomy and privacy, and initiate discussions with regard to the ethics of data collection (Schroeder, Cowls 2014). This is very important aspect of using *Big Data* in research processes. Much of the data is collected in a commercial manner and used for the performance of instrumental purposes. Social engineering is one area which becomes the beneficiary of the *Big Data* trend.

We would like to stress that growing public awareness of the security of privacy is a positive process (see Sherly case) but we would also welcome a discussion on the advantages of using *Big Data*. If we want to incorporate *Big Data* to the research process we have to focus on the following issues:

a. What are key methodological issues considering *Big Data*?

Use of *Big Data* in real-time research is a challenge considering relevance of data. So far studies of innovations have most often used *post factum* analysis.[6]

[6] Miller Carl and Bobby Duffy. *The birth of real-time research*:http://quarterly.demos.co.uk/article/issue-2/the-birth-of-real-time-research/#article-footer [access: 20.08.2014].

They have tried to capture changes in the conceptual framework, the diffusion of innovation, and the final results in the form of case studies. Therefore, the study of innovation often focuses on new technologies. In them the product's lifecycle, means of validation, and the market effect become indicators that researchers attempt to define and characterize. Modern tools have led to breakthroughs in research in real time, which in the case of the analysis of innovation is particularly important because the legitimacy of a new project is always associated with social overtones. Today, the importance of research *past*, *present* and *future* provides additional cognitive opportunity ("social life of methods" (Ruppert 2013)), but also makes the conduct of trials itself become the object of analysis ("object of the study vs study itself" (Back et al. 2013)). Innovation research using modern technology allows us to describe processes including detailed analysis of units of time, and the attributes of social interaction but also, importantly, it can be examined by several investigators at the same time, as the assumptions, conclusions, and obtained data are subject to mutual verifiability ("plurality of vantage points").

We have been observing a redefinition of the current understanding of the linearity and continuity of the research process. Points of gravity associated with the conceptualization of research, data collection and analysis are seemingly being reorganized. Modern data analysis can precede the study. It is the most important advantage of modern social research methods supported by advanced algorithms. The data should allow us to make an initial analysis before designing the research, preparation and selection of indicators, and creating hypotheses. Among the opponents of the real-time research, there is a belief that this kind of research often results in an inability to grasp social change due to too much "focus on the present." In fact, the analysis of the here and now, with the conscious collection of information at a time and their deposition in various forms (notes, databases, descriptions) allows us to understand the true nature of social change. Separation of the place and object of study seems to no longer be relevant. The argument suggesting that users of new media form a separate social group based on which we are unable to quantify the results of the research is increasingly questioned, because of the breadth and scale of the data obtained and their level of significance.

Questions about when and where we should carry out research include both classical forms of research as well as the modern computerized forms, with the difference that thanks to new technologies we can eliminate the concept of time and space and perform research on a wider scale, which a few years ago was completely impossible. Today we are able to draw the first conclusions on specific social problems in a few hours or days by analyzing qualitative and quantitative data available on the network. Findings of these studies are the subject of worldwide debate about the future of society, its development and standards and values that it manifests. An excellent example of a constructive dialogue on social research is the international project "Collaborative Online Social Media Observatory (COSMOS):

Social Media and Data Mining"[7], which is an attempt to systematize scientific inquiry of material available through open data and social media.

b. How can we effectively employ *Big Data* in research strategies?

To answer this question we have to focus on two issues: the methods or techniques used to extract data and the methods used to analyse them. As with many other dichotomies (macro–micro, quality–quantity) the strict borders between these issues are also blurred. Modern research, as well as social research, with the right knowledge can be supported by data provided to us by special tools. The crawling or scraping techniques of collecting data are at the same time pre-analytic tools which create samples which we can further research via other tools of e.g. textual or network analysis (Marres, Weltevrede 2013).

Thanks to modern technology we can conduct qualitative and quantitative research at the same time. Qualitative analyses of voluminous and diverse materials can be conducted in an automatic way. There are no synthetic studies with well described research methods but there are numerous interesting studies covering parts of *Big Data* (e.g. Tinati et al. 2014; Ampofo 2011). Studies using big data are mainly interdisciplinary research so we have to adjust to the methodological eclecticism and mostly mixed methods approaches. In building a methodology for the use of *Big Data* we would rather collect good studies as examples than develop a holistic research program.

Software that uses a simple interface allowing us to make simple but meaningful analysis is certainly helpful. In standard sociological research we have to extract a sample to create general assumptions (from micro to macro level). *Big Data* provides us with new opportunities and challenges. We have to analyse big volumes of data to prepare starting hypotheses and assumptions and after that we can focus on specific issues (so we can reverse from the macro to the micro level). Therefore, big data research is seen as a new version of positivist empirical research.

Taking the above remarks into account, the primary problem in using *Big Data* is for us not the lack of methodology itself but the problem of choosing and employing the right one. Using *Big Data* is definitely related to skills of proper analysis of extracted data, and substantive preparation of work, often with heterogeneous and redundant materials. The introduction of *Big Data* into the social sciences also requires sensitivity and sociological imagination. Sociology and sociologist do not have monopoly over collecting and analyzing data but they can provide critical approach to simplistic generalizations on digital technologies and

[7] *The Collaborative Online Social Media Observatory (COSMOS) is an Economic and Social Research Council (ESRC) strategic "Big Data" investment that brings together social, computer, political, health, statistical and mathematical scientists to study the methodological, theoretical, empirical and technical dimensions of social media data in social and policy contexts*: http://www.cs.cf.ac.uk/cosmos/ [access: 20.08.2014].

data they accumulate. This is important especially today when we have so many ways to describe reality, but the answers to the questions of why are highly limited.

c. How reliable is *Big Data*?

Big Data, in its scope and availability, is a new phenomenon, very briefly examined. As a source and method of data acquisition it still requires research in sociological contexts. It is very difficult to develop one correct approach or research procedure. Before we use the collected data, we need to verify the sources, methodologies and social groups they cover. It is extremely important to use reliable data sources, because it shapes the research process and influences final results. The data obtained from publicly available sources require additional validation and verification, as they are often the result of personal narrative embedded in a broader context. The discussion started over the methodological issues connected with mass data will be very helpful for the implementation of big data. This particularly concerns measurement and sample quality (we do not touch it in our text) as well as causes of distortions of data and possible ways of avoiding or controlling them (Baur 2009). We have to bear in mind that data is collected for different purposes and not everything is available (some databases are not available at all and some only on a commercial basis – the issue of independence, continuity and access to the IT infrastructure).

That brings into discussion the issue of the very existence of 'raw data'. Modern devices such as smart phones, wearable computers and tablets ("multi-screen world")[8], become a kind of gatekeepers of information, imposing the practice of communicating with the medium and the format of the data that we receive from them. We have to be aware that they possess 'algorithmic authority' (Rogers 2013), so we cannot treat information's they provided as pure but also as social data embedded to power relations. Thus, we are rather inclined towards the arguments that any data is already marked with conceptual categories and because of that data and analysis cannot be distinguished in any clear way, nor can the processes of data collection and data analysis be separated (Marres, Weltevrede 2013). That approach is changing the focus of using *Big Data*, from sources of data to the methods of analysing them. So, sociologists using empirical technologies measure and format phenomena according their theoretical assumptions, despite original intentions, or lack thereof, in creation of digital traces.

To prevent *Big Data* from serving only large corporations, governments and industry, sociologists should actively involve the concept and techniques of data collection in the field of their analysis. This is also relevant to the responsibility for the results of the research, understanding of the data, in terms of who creates

[8] *The new muti screen-world study research*: http://think.withgoogle.com/databoard/media/pdfs/the-new-multi-screen-world-study_research-studies.pdf [access: 22.08.2014].

it, whether it was created consciously or unconsciously, and who would like to apply it only for the purposes of social engineering. Today, the boundary between the offline world and the online world is blurred, we are living in extended reality. An excellent example of this is the paradigm of the Internet of Things, the role of non-people who surround us hidden behind the data, but act as participants in social life. From the point of view of academic analysis the real-time translation of the surrounding reality and social behavior is now of crucial value and a challenge.

Literature

Ampofo L., 2011, *The Social life of real-time social media monitoring*, "Participations: Journal of Audience & Reception Studies" 8: 21–47.

Back L., Lury C., Zimmer R., 2013, *Doing Real Time Research: Opportunities and Challenges*, Discussion Paper. NCRM: http://eprints.ncrm.ac.uk/3157/1/real_time_research.pdf [access: 22.08.2014].

Baur N., 2009, *Measurement and Selection Bias in Longitudinal Data. A Framework for Re-Opening the Discussion on Data Quality and Generalizability of Social Bookkeeping Data*, "Historical Social Research" 3: 9–50.

Baur N., 2011, *Mixing process-generated data in market sociology*, "Qual Quant" 45: 1233–1251.

Beinhocker E.D., 2007, *The Origin of Wealth. Evolution, Complexity, and Radical Remaking of Economics*, London: Random House Business Books.

Bell D., 1999, *The Coming Post-Industrial Society*, New York: Basic Books.

Bourdieu P., 2005, *The Social Structure of the Economy*, Cambridge: Polity Press.

Bughin J., Chui M., Manyika J., 2013, *Ten IT-enabled business trends for the decade ahead*, "McKinsey Quarterly", May 2013.

Castells M., 2000, *The Rise of the Network Society*, Oxford: Blackwell Publishers.

Dosi G., Orsenigo L., Labini M.S., 2005, *Technology and the Economy*. In: N.J. Smelser, R. Swedberg, eds., *The Handbook of Economic Sociology*, Princeton: Princeton University Press, 678–702.

Fligstein N., Mc Adam D., 2012, *A Theory of Fields*, Oxford: Oxford University Press.

Florida R., 2001, *The Rise of Creative Class: And How It's Transforming Work, Leisure, Community and Everyday Life*, New York: Basic Books.

GCR, 2013, *The Global Competitivness Report 2013–2014*, http://www.weforum.org/reports/global-competitiveness-report-2013-2014

GCR, 2014, *The Global Innovation Index 2014. The Human Factor in Innovation*, http://www.globalinnovationindex.org/userfiles/file/reportpdf/GII-2014-v5.pdf

Lupton D., 2015, *Digital Sociology*, London: Routledge.

Marres N., Weltevrede E., 2013, *Scraping the Social? Issues in life social research*, "Journal of Cultural Economy" 6.3: 313–335.

Ries E., 2011, *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*, New York: Crown Publishing.

Ritzer G., 2004, *The Globalization of Nothing*, London: Sage.

Rogers E.M., 2003, *Diffusion of innovations*, 5th edition, New York: The Free Press.

Rogers R., 2013, *Digital Methods*, Cambridge, MA: The MIT Press.

Ruppert E., 2013, *Rethinking Empirical Social Sciences*, "Dialogues in Human Geography" 3.3: 268–273.

Schmitt J., 2014, *Social Innovation for Business Success*, London: Springer.

Schroeder R., Cowls J., 2014, *Big Data, Ethics, and the Social Implications of Knowledge Production*, Paper presented at Data Ethics Workshop, KDD@Bloomberg, August 24 in New York, USA.
Schumpeter J.A., 2000, *Entrepreneurship as Innovation*, In: R. Swedberg, ed., *Entrepreneurship, Social Science View*, Oxford: Oxford University Press, 51–75.
Steyaert C. Hjorth D., 2006, *Introduction: what is social in social entrepreneurship?* In: C. Steyaert, D. Hjorth, eds., *Entrepreneurship as Social Change*, Cheltenham: Edward Elgar, 1–20.
Tinati R., Halford S., Carr L., Pope C., 2014, *Big Data. Methodological Challenges and Approaches for Sociological Analysis*, "Sociology" 48.4: 663–681.
Tonkiss F., 2006, *Contemporary Economic Sociology. Globalization, production, inequality*, London: Routledge.
Touraine A., 1974, *The Post-Industrial Society*, New York: Random House.
West B., Turalska M., Grigolini P., 2014, *Networks of Echoes*, London: Springer.