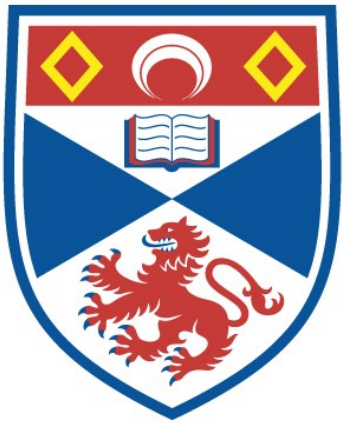
PROJECT MANAGEMENT IN SOCIAL DATA SCIENCE: INTEGRATING LESSONS FROM RESEARCH PRACTICE AND SOFTWARE ENGINEERING

Ilia Lvov

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2019

Full metadata for this item is available in
St Andrews Research Repository
at:
http://research-repository.st-andrews.ac.uk/

Please use this identifier to cite or link to this item:
http://hdl.handle.net/10023/18936

This item is protected by original copyright

Project Management in Social Data Science: integrating lessons from research practice and software engineering

Ilia Lvov

University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of

Doctor of Philosophy (PhD)

at the University of St Andrews

June 2019

Candidate's declaration

I, Ilia Lvov, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 80,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree.

I was admitted as a research student at the University of St Andrews in August 2014.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date: 12.06.2019 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date: 12.06.2019 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Ilia Lvov, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date: 12.06.2019 Signature of candidate

Date: 12.06.2019 Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Ilia Lvov, understand that by declaring that I have original research data or digital outputs, I should make every effort in meeting the University's and research funders' requirements on the deposit and sharing of research data or research digital outputs.

Date: 12.06.2019 Signature of candidate

Permission for publication of underpinning research data or digital outputs

We understand that for any original research data or digital outputs which are deposited, we are giving permission for them to be made available for use in accordance with the requirements of the University and research funders, for the time being in force.

We also understand that the title and the description will be published, and that the underpinning research data or digital outputs will be electronically accessible for use in accordance with the license specified at the point of deposit, unless exempt by award of an embargo as requested below.

The following is an agreed request by candidate and supervisor regarding the publication of underpinning research data or digital outputs:

No embargo on underpinning research data or digital outputs.

Date: 12.06.2019 Signature of candidate

Date: 12.06.2019 Signature of supervisor

Abstract

Online platforms, transaction processing systems, mobile sensors and other novel sources of data have shaped many areas of social research. The emerging discipline of social data science is subject to questions of epistemology, politics, ethics and responsibility, while the practice of doing social data science raises significant project management issues that include logistics, team communication, software system integration and stakeholder engagement. Keeping track of such a multitude of individual concerns while maintaining an overview of a social data science project as a whole is not trivial. This calls for provision of appropriate guidance for holistic project management.

The project management issues in social data science are strikingly similar to those arising in software engineering. In this thesis, I adapt a particular software engineering project management tool – the SEMAT Essence model (Jacobson et al., 2013) – to the needs of social data science. This model offers a holistic management approach by addressing key project aspects, including the often overlooked yet crucially important ones such as maintaining stakeholder engagement and establishing the ways of working. The SEMAT Essence is a progress tracking model and does not assume any specific work process, which is valuable given the great diversity of social data science projects.

To achieve this goal, I study the practice of doing social data science through participant observation of social data science projects and by providing ethnographic accounts for those. Using the ethnographic findings and the basic content and structure of the SEMAT model, I develop the Social Science Scorecard Deck – an agile project management tool for social data science. To assess the Scorecard Deck, I use the tool in management of a social data science project and then subject the tool to external validation by interviewing experts in social data science.

Acknowledgements

This work was supported by the University of St Andrews (the 7th Century Scholarship).

I would like to thank the organisations that allowed me take part in their projects and use them as case studies: The Open University, The University of Edinburgh and The Wire Free Production company. A special thanks goes to the British Council for reviewing and approving the Shakespeare Lives evaluation case study. I would like to personally thank Dr Angus Bancroft for agreeing to test-drive the Social Data Science Scorecard Deck. Finally, I would like to express my gratitude to all the interview participants.

The digital outputs underpinning this thesis (the Social Data Science Scorecard Deck) can be accessed at:

https://doi.org/10.17630/880c2780-120c-48f8-9ccb-aff048397d57

CONTENTS

List of Figures

List of Tables

CHAPTER ONE

INTRODUCTION

The world of data is going through fundamental changes that concern the sorts of data that are acquired, the methods and technologies that are applied to them and even the subject of knowledge that results from these processes (boyd and Crawford, 2012). This can be colourfully illustrated by the appearance of a range of new terms. Just a few examples are "big data" (Jagadish et al., 2014), "open data" (Janssen et al., 2012) and "data science/scientist" (Davenport and Patil, 2012; Dhar, 2013) that captures a notion that data become a subject of expertise on their own. As Tauberer (2014) argues, while one may try to dismiss those terms as mere buzzwords, they aid in navigation through the emerging trends in the world of data and indicate changes in our perception of data, growing interest in them and rising expectations of them.

The changes in the world of data have a potential to provide new insights into the life of societies and their members. In the academic environment, they allow posing and answering new questions in social science disciplines. For example, Kitchin (2013) discusses the use of new forms of data in human geography. Mao et al. (2011) and many others apply computationally intensive methods to new forms of social data to forecast changes in financial markets. Many actors outside of academia seek knowledge about people as well – businesses are looking for marketing insights (Erevelles et al., 2016), journalists produce independent investigations of societal issues with data-driven methods and seek to present their findings in forms digestible for wider audiences (Gray et al., 2012), governments employ data for the needs of national security (Kim et al., 2014), while civic movements, in turn, demand openness of data from the government (Janssen et al., 2012) and promote unconventional use of these data to improve peoples' lives (Baraniuk, 2013). Moreover, as this thesis will later show, new forms of collaboration on the intersection of academic- and non-academic settings are also emerging.

The scope of the social issues tackled with data-driven methods also varies greatly. For example, a study by Servia-Rodríguez et al. (2017) looks into the well-being of individuals. Many studies

look at very specific social groups and populations. For example, much of recent research studies the behaviour of users of a particular social media platform, with Twitter being the most frequently studied (e.g. Zubiaga et al., 2016). That being said, some studies are in fact concerned with questions of society as a whole. For example, Burrows and Savage (2014, p. 3) report on the experience of the Great British Class Survey – a "a hybrid project, which spliced together a fairly conventional social survey (providing accounts of actions), with a high-profile web platform hosted by the BBC asking a battery of questions about respondents' economic, social and cultural capital".

Quite often the data employed in the new forms of social research satisfy some quantitative definition of "bigness" – for example, the traditional 3Vs of volume, variety and velocity or any of their extensions (cf. Laney, 2001; Ari et al., 2012; Marz and Warren, 2015; Kitchin, 2013) – however, this is not necessarily the case. For example, Psylla et al. (2017), in their study of gender's impact on social interactions, construct graphs of Facebook interactions for a limited sample of students that are subsequently contextualised with other data types, including traditional survey data. The resulting data are by no means big, but the process of their construction (i.e. systematically gathering records of Facebook interactions, transforming them into a meaningful dataset and putting the data into context of data from other sources) requires actions and decisions uncommon for traditional social research.

The example above shows how the new forms of research embrace the emerging world of digital traces that allow us to navigate back and forth between detailed and aggregated descriptions of social processes (Latour, 2009; Procter et al., 2015). Some of these traces may be transformed into data akin to the naturally occurring data known to traditional social sciences – e.g., some of social media user-generated content is similar to dialogue transcripts studied by sociologists engaged in conversation analysis (Wetherell, 1998) and some transnational data is familiar to economics and management (cf. Hallowell, 1996; McInish and Wood, 1992). Other forms of digital traces (e.g. sensor data) are out of scope of traditional social sciences (a detailed taxonomy suggesting 26 types of new social data can be found in Kitchin and McArdle, 2016). The value of data resulting from different digital traces can thus be maximised by interlinking and aggregation (boyd and Crawford, 2012).

At the same time, as Leek (2013) points out, the issues of data sources and tools involved in data-intensive research activities often wrongly out-shine those of adhering to the *scientific ethos*, which Leek operationalises as making sure that the available data and methods are a valid means for answering the *research questions*. Kakati (2017) shares this sentiment and focuses his criticism on frequent misuse of statistical methods in data-driven studies. Yet, the complexity of *science* in data-intensive research transcends mere issues of applying the statistical apparatus – it

gives a new spin to key issues in epistemology (Kitchin, 2014; Burns, 2015). It may not only be difficult to judge whether the research findings are valid – it is sometimes hard to say whether the research questions have been correctly posed in the first place. This suggests that the crucial aspect of scientific ethos in data-driven research should be *continuous maintenance of critical, self-reflective stance towards own research activities*, which involves *keeping track of- and accounting for the decisions and operationalisations involved, understanding their limitations and assessing the validity of the process as a whole and of the derived findings*. In the world of the ubiquitous epistemological crisis that goes way beyond research inquiry[1], data-intensive research, as a new paradigm of knowledge discovery (Tansley et al., 2009), simply cannot afford to ignore these issues.

1.1 Social Data Science Definition

Reflecting on the discussion of the new opportunities for studying the social, the new actors emerging in this the field and the new challenges of adhering to the scientific ethos, I would like to suggest *social data science* as the key term for this thesis and to provide the following definition:

> *Social data science (SDS)* is the whole range of activities, both academic and non-academic, that seek to gain verifiable and robust actionable knowledge about societies, their groups and individual members through embracing a variety of forms and types of data and investigating them within a self-reflective research process using well-informed methods.

This term, while in use in the literature, is not the most popular one among those that aim to capture the essence of new form of social inquiry. One may quite often see it as a publication keyword (cf. Margetts, 2016; Gao et al., 2017), but there have been only few attempts to thoroughly discuss it. Therefore, I would like to motivate my choice of terminology.

First of all, the parental term *data science* is certainly among the most popular ones. Burrows and Savage (2014) show that "data scientist" has become a more frequent search term on Google than "statistician" in 2013. At the moment of writing, the frequency of searching for "data scientist" is at a historic peak with approximately 5 times more searches per month than for "statistician"[2].

[1] One example of this crisis is the ongoing discussion of critical thinking and media literacy, cf. boyd (2018); Doctrow (2018).

[2] `https://trends.google.com/trends/explore?date=2005-01-01%202018-04-27&q=statistician,Data%20Scientist`. Accessed on 2018-04-27.

This off-the-chart popularity has led Davenport and Patil (2012) to famously pronounce data scientist "the sexiest job of the 21st century". Cassel et al. (2017) argue that the basics of data science should be taught to undergraduate students of all disciplines.

Second, *social* data science is not an obscure term by itself either. Arguably, it is most widely recognised within the United Kingdom. For example, multiple research initiatives unequivocally specialise in social data science. Cardiff University hosts the Social Data Science Lab[3]. The Alan Turing Institute – a national cross-university research institution in data science – has a Social Data Science interest group[4]. One of its participating academic bodies, Oxford Internet Institute of the University of Oxford, even offers an MSc in Social Data Science[5]. The international recognition of social data science as a valid term may be illustrated by the University of Copenhagen's Centre for Social Data Science[6] and the Social Data Science lab in Montréal [7] – this is not to mention that other organisations may recognise and adopt the term without using it as part of their title.

The most important reason for choosing the term "social data science" is its ability to capture the essence of this thesis' subject of inquiry and being a natural match for the suggested definition. A particular aspect of this definition that is worth mentioning in this regard it the use of the phrase "actionable knowledge". Indeed, this phrase is something of a tautology: according to the widely accepted Knowledge Hierarchy model (Rowley, 2007), knowledge is by definition transferable into instructions for further actions. Yet, knowledge's potential for action is worth re-emphasising to strengthen the notion that social data science has direct *impact* on human lives and society in general and thus making sure that social data science is done in accordance with good research principles and practice is of *significant societal importance*.

1.2 Research Design and Thesis Structure

In this thesis, I argue that the complexity of doing social data science transcends the questions of methodology and raises significant project management issues that include logistics, team communication, software system integration, stakeholder engagement and ensuring compliance with ethical and regulatory frameworks. Keeping track of such a multitude of individual aspects while maintaining an overview of the project as a whole and ensuring a *shared understanding* of the key project's decisions and operationalisations within the team is far from trivial. This calls

[3] http://socialdatalab.net/
[4] https://www.turing.ac.uk/research_projects/social-data-science/
[5] https://www.oii.ox.ac.uk/study/msc-in-social-data-science/
[6] http://sodas.ku.dk/
[7] https://socialdatasciencelab.org/

for provision of *guidance for holistic project management of social data science projects*.

The project management issues in social data science are strikingly similar to those arising in software engineering. Through decades of existence, software engineering (Sommerville, 2016) has gone through a number of shifts to increasingly mature project management paradigms that have led to production of numerous pieces of increasingly robust management guidance. Given this, I propose a **thesis** that *there is significant value for social data science in adapting the guidance from software engineering for the needs of holistic project management* and design my research to address the corresponding **goal** of *implementing a successful adaptation of such guidance*.

To progress towards this goal, I set out to achieve the following **objectives**:

1. to contextualise my work in understanding of the grand challenges that social data science faces as a discipline;

2. to make an informed decision on what project management guidance from the field of software engineering can be effectively adapted to social data science;

3. to discover the particular issues that arise in the practice of doing social data science, how those issues intertwine and amplify each other and how they are dealt with by social data science project teams;

4. combining the outcomes of objectives 1-3, to implement a tool for holistic project management in social data science;

5. to use the implemented tool in the management of a social data science project;

6. to subject the tool to external validation.

The main body of the thesis is structured in accordance with these objectives. Chapter 2 addresses the objectives 1 and 2 through **reviewing literature** on challenges in social data science (see Sections 2.1, 2.2 and 2.3) and on relevant project management guidance (see Section 2.4). Chapter 3 addresses objective 3 through reporting on **ethnographic case studies** informed by **participant observation** of social data science projects. The ethnographies produced stand as a significant contribution on their own. Chapter 4 addresses objective 4 by discussing the processes of **systematically informing the design** of a project management tool for social data science – the Social Data Science Scorecard Deck. Chapter 5 addresses objective 5 by reporting on an **in-depth case study evaluation** of the Social Data Science Scorecard Deck (see Section 5.2)

and objective 6 by providing an external evaluation of the Social Data Science Scorecard Deck through **interviews with experts in social data science** (see Section 5.3).

CHAPTER TWO

KEY CHALLENGES IN SOCIAL DATA SCIENCE: LITERATURE REVIEW

The Introduction has provided the definition for social data science and suggested that it is a highly problematic field of activity. It is thus logical to proceed to a critical analysis of the major known issues in social data science and to examine how a project management tool that captures the lessons from software engineering may assist in coping with these issues. Given the *social* nature of the discipline, I choose to start with discussion of the problematic relationships between social data science and society.

2.1 Social Data Science and Society

In the following discussion of the relationship between social data science and society I will draw heavily on literature in science and technology studies (STS). Indeed, as a discipline that relies on data generated through new technologies and on methods that are implemented in new technologies, social data science goes hand in hand with technological development and innovation (Edwards et al., 2013; Halfpenny and Procter, 2015; Wright, 2014; Jagadish et al., 2014). The STS frameworks can help to assess the roles of different social groups in the development of tools and methods for social data science and to discuss the social and political implications of this division of roles.

2.1.1 Social Shaping of Supporting Technologies

As outlined above, social data science is continuously supported by technological development. These developments do not happen in vacuum – rather, technology and society continuously interact and *shape* each other (Williams and Edge, 1996). Therefore, social factors might influence the prioritisation with which different developments that support social data science are produced. As such, one important factor for this prioritisation is what *perceived advantage* a potential technological development is expected to bring. Such advantages may, for example, be managerial (McAfee et al., 2012), political (Janssen et al., 2012) or scientific (Raghavan, 2014). MacKenzie and Wajcman (1999) point out that specifically the economic or business advantage is commonly the strongest motivation force. Moreover, the technological developments for social data science are likely to happen in the areas where growing expertise in dealing with new forms of data leads to increasing returns on adoption (cf. Arthur, 1994) of data science.

The notions above correlate with the history of innovations that support social data science. In its early days, social data science largely relied on so-called "big data" technologies that developed in- and were adopted by large hi-tech companies. For example, the famous MapReduce computational paradigm for parallel data processing was designed at Google (Dean and Ghemawat, 2008). It has subsequently received an open-source implementation in Apache Hadoop whose co-founder Doug Cutting was Yahoo's employee at the time and led the development project in this capacity (Shvachko et al., 2010). Many of the most famous earliest methodological innovations in social data science also came from the hi-tech industry – for example, collaborative filtering algorithms for personalisation of recommendations (Linden et al., 2003; Das et al., 2007) and Google's PageRank algorithm for ranking importance of nodes in larges-scale networks of linked texts (Page et al., 1999).

In the recent years, the dominance of hi-tech companies specifically in *social* data science has started to decrease. First of all, the cutting-edge technical and methodological developments in data science developments seem to have shifted from studying and predicting human behaviour, traits and preferences (e.g. from the "social" side) to progressing the development of artificial intelligence with the help of deep learning, for example, in the form of advancing speech and image recognition techniques (LeCun et al., 2015; Dahl et al., 2012; Mordvintsev et al., 2015). This gives a chance for later adopters (cf. Rogers, 1995) of social data science to "catch up" and to critically reflect on how they can bring their experience in traditional forms of empirical social research to new realities (Levallois et al., 2013). Appearance of new open-sourced computational frameworks such as Apache Storm and Apache Spark (Oliver, 2014) is also a liberating change – although those frameworks are still mostly oriented at real-time business intelligence (Apache Foundation, nd; Oliver, 2014) in alignment with recommendations of Davenport et al. (2012),

they are more light-weight than Hadoop, less rigid in terms of suggested data processing pipeline, provide convenient APIs for many programming languages and thus can be repurposed with more flexibility.

What is arguably even more important for wider adoption of social data science is the appearance of technologies specifically for dealing with datasets whose complexity is higher than supported by standard data analysis software such as spreadsheet software yet not high enough to require a complex multi-node hardware infrastructure. Voss et al. (2017) refer to such datasets as "small big data". One group of such technologies facilitate parallel computations on a single multi-core machine. For example, the Streams API for parallel computing was added to Java in version 8 of the language (Urma et al., 2015). Python and R – Programming languages of ever increasing popularity in the field of data analysis (Robinson, 2017) – also start supporting parallelism through dedicated modules and packages (Raschka, 2014; Bengtsson, 2018). Utilising the computational power of a single workstation or server is relatively straightforward compared to technologies like Hadoop (and even Spark, cf. Zaharia et al., 2010) that are aimed at performing computations in connected clusters of multiple computers. While Hadoop and Spark can run on a single machine, their overheads would be disproportionately high compared to the infrastructure complexity – and starting up a cluster is often neither feasible nor required (Stucchio, 2013).

Thanks to technological and methodological advancements, social data science has started to be adopted by a diverse set of social actors such as journalists (Gray et al., 2012), civic movements (Townsend, 2013), smaller businesses (Coleman et al., 2016) and authorities (Berk et al., 2016). However, this does not by itself bring true democratisation to social data science due to the inequalities in access to- and control of the new forms of social data themselves, which by themselves become a valuable resource of power. The next section will look at these issues through the prism of politics of data science.

2.1.2 Politics of Social Data Science

Social data science does not simply provide new tools and methods to gather knowledge for its own sake. It is used by influential social actors, supports their actions and informs their decisions, and as such it shapes society. For this reason, Green (2018) calls for data scientists to consider themselves *political actors*. This idea of social data science being politically non-neutral follows propositions of politics of technology (Winner, 1980) and the politics of method (Steinmetz, 2005; Savage, 2010). Technologies and methods, according to these schools of though, can serve political agendas and shape the distribution of power and authority in the society – and social data science exists on an intersection of technologies and methods.

Data by themselves are also becoming increasingly political, as varying levels of data access and control create digital divides in the society (boyd and Crawford, 2012; Savage and Burrows, 2007) and raise questions of ownership of data, human rights, and ethics and privacy (Savage and Burrows, 2007; Ruppert, 2013). The distribution of access to new forms of social data is strongly skewed in favour of a limited number of big corporations (Törnberg and Törnberg, 2018; Department for Digital, Culture, Media and Sport, 2018a). This inequality in access to data does not only mean that other actors in social data science have a hard time competing with the big players; it also raises questions of whether the owners of large datasets are effectively capable of public surveillance (Fuchs, 2017, pp. 52-61, 183-207), or at least of knowing more than the data they own immediately suggest. Indeed, social data science methods allow to use the digital traces (Latour, 2009) that individuals routinely leave in the form of user-generated-, transaction- or sensor data to draw reliable inferences about private aspects of human life that individuals do not directly disclose (Kosinski et al., 2013) and to link disparate anonymised data records with a person's identity (De Montjoye et al., 2015).

The example of using social data science as a means of gaining power unfairly that has arguably sparked the biggest public and journalistic outcry thus far is the one related to manipulation of voter opinions before the US Presidential Election (Rosenberg et al., 2018; Cadwalladr and Graham-Harrison, 2018) and the Brexit referendum (Cadwalladr, 2017) in 2016. The "Leave" referendum campaign and the Trump's presidential campaign were accused of using Facebook data for micro-targeting, i.e. for selecting individual recipients of promotional messages and for tailoring the message to the exact social, economic, demographic and psychological characteristics of each recipient or of small groups (Papakyriakopoulos et al., 2018).

Micro-targeting was not by itself a new technique. It had been known in marketing for almost a decade (cf. Agan, 2007). In the context of politics, its aspects were employed in Obama's presidential campaign in 2008 (Franz and Ridout, 2010). The ethical concerns of micro-targeting were also not new, as they started to appear almost simultaneously with the initial spread of the technique (Barbu, 2014). Korolova (2010) questioned the ethical and privacy considerations of using social media data for micro-targeting.

New to the 2016 scandals were the *data breaches* (Cadwalladr and Graham-Harrison, 2018): Cambridge Analytica – a data analytics company that provided consultancy for the accused campaigns – accessed data on at least 50 millions of Facebook users "without meaningful consent" (Green, 2018, p. 3) and used them against the platform terms and conditions. Lack of action from Facebook to prevent Cambridge Analytica from misusing the data further amplified the public critique, as Facebook's own data-driven algorithms also likely contributed to Trump's victory, albeit most probably unintentionally. The company's algorithm that compiled the "Facebook

Trending" feature – a chart of the most popular news stories – did not manage to filter out unreliable news entries, or "fake news" (Thielman, 2016). While there were political fake news that supported both Trump and his opponent Clinton, the frequency of the pro-Trump fake news was much higher (Allcott and Gentzkow, 2017). Thanks to Facebook Trending, the already high reach of the those articles was further increased.

Interestingly, the topic of the 2016 US presidential elections can also show how social data science can potentially empower society. Indeed, it provides an example of social data science being used to oppose another instrument of public opinion manipulation – social media bots that spread promotional messages while pretending to be human users. Bessi and Ferrara (2016) use contemporary data analysis techniques to compare tweets by bots and genuine users who support either Trump or Clinton. The findings of this study may equip those who read political debates on Twitter with better understanding of how bots behave, so that they could spot bot tweets in the conversations. To construct samples of bot and human post, Bessi and Ferrara (*ibid.*) employ "BotOrNot" (Davis et al., 2016) – a bot-detection system that also employs data-driven algorithms. Similar techniques are applied by Narayanan et al. (2017) to Twitter posts around Brexit. Such studies demonstrate the potential social value of social data science. The juxtaposition of these studies with the scandals surrounding Facebook, in turn, raise the questions of *responsibility and accountability* in social data science.

2.1.3 Accountability and Responsibility in Social Data Science

As social data science may equally bring value to the society and, when misused, pose societal risks, it is important to consider how the notions of responsibility and accountability can be incorporated into its practice. Given the strong ties between social data science and methodological and technical innovation, a natural first step in that is examination of the *responsible innovation* principles. Responsible innovation is a relatively new and broad concept. While its precise scope is a bit ill-defined, its widely cited definition is given by von Schomberg (2011, p. 50):

> "A transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society)."

From a historical perspective, the concept of responsible innovation can be perceived as a

successor of various earlier concepts, including "anticipatory governance, technology assessment, and upstream public engagement" (van Oudheusden, 2014, p. 68). The constructive technology assessment school of thought seems to be of the most direct relevance, at it argues for assessing technologies not through observing their impact, but through predicting it in advance (Schot and Rip, 1997) and can help ensure that innovations meet existing societal demands (Genus, 2006; van Merkerk and Smits, 2008). This notion is closely related to the public engagement with science movement, which also considers responsible innovation to be its successor (Stilgoe et al., 2014).

Responsible innovation poses the questions of who exactly is responsible, how they are responsible and for what (Grinbaum and Groves, 2013; Stahl et al., 2013). Because of this, the concept can be operationalised in various ways. For example, Stahl et al. *(ibid.)* attempt to provide a framework for responsible and ethical developments in ICT. They identify the ethical issues associated with emerging types of ICTs and suggest some specific institutional measures to address those such as provision of ethical regulation for ICTs. A more general framework for responsible innovation is suggested by Stilgoe et al. (2013). They suggest that responsible innovation should be:

- *anticipative*: consequences of the developments should be predicted and evaluated in advance;
- *reflexive*: the steps in the developments that are already performed should be subject to questioning and critique by the developers;
- *inclusive*: the developments should be public-oriented and open to public discussions and engagement;
- *responsive*: the critique should bring practical actions.

This framework focuses on the core properties of responsible innovation itself rather than some particular associated issues and measures. This makes it useful as a generic starting point for tailoring to specific disciplines. Below I attempt to do that by formulating a list of *social data science responsibility principles*.

2.1.3.1 Social Data Science Responsibility Principles

Following the lines of Stilgoe et al. (2013), I do not attempt at providing a list of specific measures that should be taken in a social data science project to ensure responsible approach to research – rather, I list some of the characteristics that social data science needs to satisfy to be ethical, considerate of societal- and stakeholder interests and, overall, responsible:

Being considerate of biases coming from data, methods and researchers. Responsible data science should be aware of possible imperfections in data. This is especially relevant in the modern era of "big data'. As boyd and Crawford (2012) put it, "big" does not mean "good" if data fail to represent relevant characteristics of the studied phenomenon. Similarly, regardless of how complex and flexible the methods of data analysis are, they bring along their own assumptions and hence biases. An example of this is a study by Cannarella and Spechler (2014), who use a complex data-driven methodology to predict the changes in the number of active Facebook users. While arguably the most staggering result of their work is the prediction of Facebook's eventual decline, this outcome, to their own admission, is a property of the underlying statistical model and not of the empirical evidence. Finally, and especially important for reasoning on socially sensitive subjects such as predictive policing (cf. Seo et al., 2018), there is a great room for a researcher to bring their own biases and assumptions into social data science.

Supporting iterative and reflexive use of analytical modes. While the importance of iterative and reflective use of methods in data science in general – i.e. moving iteratively between evidence and theory and using results derived with one method to inform the use of another method – will be discussed in detail later (see Section 2.2.1), it is worth pointing out its specific importance for social data science from the perspective of responsibility. Iterative reasoning motivates deeper reflective interpretations of what has been discovered. As an example from my prior research work in social media analysis (Hutchings et al., 2015; Procter et al., 2015), the meaning behind the quantitative figures for the most cited users on Twitter becomes much clearer when one uses these figures to qualitatively study specific tweets by those users.

Being aware of performativity. As discussed above, research into social matters may impact the reality studied. While this is expected of normative recommendations that can be derived from such research, this may be an unexpected side effect of the positively stated findings (the ones that suggest that something *does* happen – not that something *ought to* happen). MacKenzie (2008) calls the latter effect "performativity". He discusses how the work of financial theorists influenced the market actors whose behaviour became better aligned with the positive model predictions. Conversely, sometimes a discovery of a pattern in social behaviour may destroy it. An example of that will be discussed later in this literature review in regard to Phillip's curve (see Section 2.2.1.1). While it may be hard to anticipate in advance whether either of the possible perfomativity effects may take place, social researchers (including social data scientists) may wish to try to at least anticipate what consequences those effects might bring.

Encouraging dialogue. As responsible data science is aiming at reflexivity and inclusion, it is beneficial to do it in interdisciplinary teams that closely collaborate together. From the experience

of my prior work, the results of technical data analysis can be very effectively interpreted and contextualised within the research questions if the analysis steps are discussed and reflected on by a wider team than includes non-technical specialists (cf. Hutchings et al., 2015). Moreover, providing stakeholders outside a project team with some modes of feedback and critique can also be of great aid in doing social data science responsibly. Taylor et al. (2014) show that asking the wider public about requirements for a data-driven project poses many challenges but also greatly informs the process.

Supporting accountability and transparency. The implication of the above is the importance of *accountability* in social data science, which first and foremost means that a project team can answer why certain decisions have been made and actions have been taken – and that they actually provide such answers on a systematic basis (Schedler, 1999). The question of whom a social data science team is accountable to and what for is inevitable in this context *(ibid.)*. While the precise answer depends on a project, it is worth pointing out that project team members share both external accountability – i.e. to the stakeholders – and internal accountability – i.e. to each other (Fuhrman, 1999). This relates to the notions of *research transparency* (Miguel et al., 2014), *reproducibility* (De Roure et al., 2011; Plesser, 2018; Gent, 2013) – and, most importantly for this thesis, to the importance of *shared understanding* within the research team.

As can be seen, ensuring that a social data science project is executed responsibly and supports accountability is by itself not trivial. The following section will examine how the epistemological issues of social data science pose an additional layer of difficulty to this problematic endeavour.

2.2 Social Data Science and Epistemology

boyd and Crawford (2012) point out that the recent and ongoing changes in the realm of data and of their use have been shaping the very perception of what knowledge is. This suggests that social data science poses significant epistemological challenges beyond those inherited from the individual disciplines that inform it. Some of these challenges come from the nature of the employed data (more on those in Section 2.2.2). Other challenges arguably transcend those associated with data. To see why this is the case, it is useful to take a principal stance that social data science, as much as research in general, derives (or at least strives to derive) knowledge through applying *research methods* to answer *research questions*. Therefore, if there is a tension between questions and methods in social data science, the notion of derived knowledge becomes particularly problematic.

2.2.1 Tension between Research Questions and Methods

As shown in the Introduction, social data science is an umbrella term for a variety of research areas. Hence, it is not entirely fair to speak about the types of questions and methods in social data science in general. Yet, it seems plausible to argue that there is a natural tendency in social data science to bring together *questions pertinent to a social- or human study disciplines* – be that sociology, economics, psychology, HCI, marketing, etc. – and the *methods of pattern recognition in high-dimensional data* – whether they go by the name of data mining, machine learning, artificial intelligence and so forth. These types of questions and methods are in tension with each other. What this tension exactly is will be examined below, but first it is worth looking at what it *is not* to contest common misconceptions.

The tension between the questions and methods of social data science should not be cast as one between "quals" and "quants". Quantitative methods have a long tradition of being used in social- and human studies. The field of economics went through "a phase of intense mathematization" (Debreu, 1991, p.1) early in the second half of the 20th century. The mathematisation concerned both the modes of communicating the economic theories (*ibid.*) and, thanks to the advances in econometrics, validation of the theories' empirical plausibility (Eichenbaum, 1995, p. 1609). Other fields of social enquiry do not necessarily employ as much mathematical apparatus as a means for theoretical modelling; however, the use of quantitative methods in empirical work is rather routine. Data produced in experiments in psychology (and, by extension, HCI, cf. Carroll, 1997) are most often subjected to statistical analysis (Ferguson, 1971). Even in sociology, the textbooks on empirical research would often make a point on covering both quantitative and qualitative methods (Neuman, 2014).

Moreover, the tension does not lie between hypothesis-driven versus data-driven approach to knowledge discovery – or, as it is often recast, between deductive (cf. Popper, 1959) and inductive (cf. Bacon, 1620) reasoning. The existence of this tension might be *partially* true for hard sciences, where deductive reasoning (effectively operationalised as *first formulate a hypothesis and then test it with data*, cf. Oldroyd, 1986) "appears to enjoy preferred philosophical status" (Kell and Oliver, 2004, p. 100), while the process of inductively finding knowledge through observing patterns in data is often seen as not robust. Indeed, by the laws of logic, we can assert effect (data) from cause (the process that generates those data and is captured by theory), but not vice versa (cf. problem of induction, Vickers, 2014).

Inductive reasoning is common in social sciences, where research (especially qualitative) is often aimed specifically at discovering hypotheses finding explanations *from* data Ritchie and Ormston (2014). Besides, even in hard science, the inductive-deductive dichotomy is a misconception.

First, hypotheses do not come from nowhere. While they can come from pure reasoning, e.g. by theory synthesis (Whitehead et al., 2016; Tygart, 1988), they also do come from new observations, i.e. through induction. In fact, the importance of iterative and reflective reasoning – i.e. moving between evidence and theory and using results derived with one method to inform the use of another method – is emphasised by many authors (Kell and Oliver, 2004; boyd and Crawford, 2012; Halfpenny and Procter, 2015). Second, as Gigerenzer (2004) points out in his critique of the "null ritual"[1], the statistical techniques (including hypothesis testing) are all different attempts to address – not resolve – the problem of induction, which inevitably underlines all science (p. 604). In fact, the null ritual, so popular among the purists of deductive reasoning, seeks accepting non-specific (in statistical terminology, *alternative*) hypotheses that are often as vague as *something is not equal to zero*. Such hypotheses are not falsifiable and thus "unscientific" (Popper, 1959).

The real epistemological tension in social data science comes from the difficulty of interpreting the patterns discovered in large multidimensional datasets in terms of the underlying social phenomena – and it is those phenomena that the questions inherited from traditional social research are interested in (cf. Ritchie and Ormston, 2014). Even if it is possible in principle, in practice the models that result from data-intensive social research are steadily becoming black-boxed for all practical purposes with the growth of computational powers and the complexity of the analysis methods. This is especially true in case when the research methods involve translating the data in human-understandable form to feature sets with features of limited interpretability (e.g. in natural language processing) or when the features in data represent proxies for the underlying phenomena rather than their direct measurements (more on that in Section 2.2.2).

Being fundamental to social data science, the tension between questions and methods is met by almost everyone who does it. Therefore, either consciously or unwittingly, social data scientists make attempts at resolving this tension between questions and methods. Those attempts normally go into one of the two principle directions – either through re-purposing the methods or through adjusting the questions. As reader may imagine, both options are problematic. In the following sections, I will discuss some of the more common modes of social data science that show the associated problems.

[1] Gigerenzer (2004) defines null ritual as "1. Set up a statistical null hypothesis of 'no mean difference' or 'zero correlation.' Don't specify the predictions of your research hypothesis or of any alternative substantive hypotheses. 2. Use 5% as a convention for rejecting the null. If significant, accept your research hypothesis. Report the result as $p < 0.05$, $p < 0.01$, or $p < 0.001$ (whichever comes next to the obtained p-value). 3. Always perform this procedure." (p. 587).

2.2.1.1 Predictive social data science

"Predictive" social data science poses questions that quite severely differ from the traditional ones. This mode is arguably the most closely associated with the notion of Big Data. It is based around a central argument that the questions of causality – and thus of theorising and explanation – are not relevant as long as correlations are examined sufficiently well for the needs of prediction (Anderson, 2008; Mayer-Schönberger and Cukier, 2013). This line of reasoning is akin the positivist outlook at research and has been historically contested in various context, especially within social studies (McIntyre and Rosenberg, 2017). The revival of this idea is arguably provoked by the dominance of data science as done by business. As businesses often use data applications for classification and prediction (Davenport et al., 2012; McAfee et al., 2012), they may feel relaxed about causality. This view is supported by Lowrie (2017), who provides the following quote from an interview with a data scientist:

> "It's an area where it's hard to separate out whether you are doing something in the name of doing good science, or in the name of doing good business, because they come together." (p. 6)

At least three major problems arise from this stance. Two are related to non-neutrality of social data science (see Section 2.1) and of responsibility and accountability in this field (see Section 2.1.3). First, if predictive algorithms that emerge out of doing social data science – or rather decisions based on those predictions – can profoundly affect society and its members, it may be crucial to understand the "why" behind those decisions. Otherwise, a decision-maker who uses a prediction system risks to not anticipate the systematic prediction errors until many false judgements are made (see discussions of systematic errors in the algorithms used in criminal justice systems, e.g. Kirkpatrick, 2016). Besides, it becomes harder to assign responsibility for unjust decisions for any particular actor (Inagaki and Sheridan, 2012).

Second, while considering only correlations and not causality may be sufficient for prediction tasks, this is not the case for planning interventions. History does know examples of interventions based solely on empirical correlations to fail dramatically. Arguably the most famous example of this predates social data science by decades. In the late 1950s a negative relationship between a country's unemployment rate and its inflation rate – the Phillips Curve – was empirically discovered (Phillips, 1958). The US government sought to use this empirical rule to contain unemployment by increasing the monetary base (printing money). This was done in spite of the theoretical explanation of the phenomenon by Friedman (1968); that explanation, among other things, suggested that the effect would wear off in the long run since economic actors would learn to *anticipate* high inflation and differentiate the associated changes in salary levels

from real increase in wealth. In line with Friedman's prediction, the USA experienced a major economic crisis accompanied by simultaneous high unemployment and high inflation in 1973-75 (McNees, 1978).

The other problem with focusing social data science solely on prediction tasks is very different in nature and has to do with the role of academic research in the face of business orientation of predictive social data science. Törnberg and Törnberg (2018) argue that the researchers still "[at least should] seek *explanation* and *understanding*" in contrast to corporate enterprises that "seek *prediction* and *control*" (p. 9, italics preserved). It is thus hardly surprising that in the field of predictive social data science the focus of academic activity among scholars with technical background shifts from substantive research on social phenomena to evaluation and advancement of the predictive methods themselves (Lowrie, 2017). This observation is an interesting counterpoint to earlier suggestion by Savage and Burrows (2007) that the advancements in academic sociology are more likely to be achieved through focusing on the social implications of the emerging empirical methods. While both issues of accuracy and applicability of the predictive methods and of their politics are of undoubted interest, I suggest that there is a moral argument to be made in favour of having a place in academic research for the more substantive social matters.

2.2.1.2 Exploratory social data science

In contrast to what predictive social data science seeks, "exploratory" social data science attempts to apply novel methods to traditional research questions. It tries to employ pattern-discovery methods and other methods applicable to high-dimensional datasets to study data *until findings that can be interpreted in terms of the underlying social phenomena are derived*. In that, they attempt to arrive at new knowledge basing on a minimum number of prior assumptions and expectations of what variable relationships would be discovered in the data.

There are several issues arising from this approach. First, for large datasets it is not always clear how to approach their exploration. In accordance with a prediction by Meehl (1978) that has empirically confirmed by Waller (2004), if a dataset contains a sufficiently large number of data points and is gathered through non-experimental research design, a difference in between-group means for *any* numerical variable in two samples formed by splitting the dataset by *any* categorical variable will mostly likely be shown to be significant by statistical testing. The same logic can be applied to all relationships in data that a researcher may look at when they first explore a new dataset, e.g. to correlations between any two numerical variables. In cases of high dimensionality – and thus a large number of variable pairs to look at – it is easy to get lost in the forest of all "significant" relationships.

This abundance of choices for how to start tackling a large dataset is often compounded by a lack of prior knowledge about the subject matter. While the advantage of inductive exploration is precisely in its ability to discover aspects of phenomena that have not been thought of before, a researcher's prior knowledge and beliefs inevitably affect what patterns they may notice in data in the first place (Silverman, 2011; Ritchie and Ormston, 2014). Unfortunately, those prior beliefs are not always sufficiently informed either because of insufficient interdisciplinarity of research teams and thus a lack of subject expertise or because the subject matters are novel by themselves. For example, Ruths and Pfeffer (2014) show why prior concepts of social relationships do not perfectly apply to user interactions on social media platforms.

As a result, while doing exploratory social data science requires considerable effort and often leads to findings that are impressive in "statistical" terms, their substantive interpretation sometimes appears to be somewhat underwhelming. For example, Stopczynski et al. (2014) provides an overview of key findings from a number of exploratory social data science projects that all relied on data collected from mobile devices belonging to large samples of a human population. For some of the cited studies, their key findings are as follows:

- People exhibit strong patterns in day-to-day mobility with a small number of frequented locations (data on 100 thousand people studied).
- Individuals tend to maintain active communication with a limited number of close social contacts (data on 20 million people studied).
- People with greater network of contacts tend to be presented with more economic opportunities (data on 65 million people studied).

Arguably, all three findings are expected with the first one being on a verge of obviousness since most people have to travel to their place of work or study from their home on a daily basis. If anything, such findings manage to show that the employed data collection methods do not lead to data that suggest false findings. But then again, a method of studying human mobility arguably needs to be off by a very large margin not to notice the existence of day-to-day patterns, and just showing that the method can pick up those is not a strong proof of its reliability. Overall, while there is value in confirming common knowledge and intuition with data, arguably there is a mismatch between the scales of the underlying datasets and the substantive significance of the findings.

2.2.2 Data Sources and Naturally Occurring Data

Another set of epistemological issues result from social data science's reliance on data that are not generated for the research purposes and within a controlled research process but occur naturally as a by-product of human activities. Such naturally occurring data raise multiple concerns as they are not tailored to a study's research questions and are often not controlled by a researcher. Even though social sciences have a long tradition of dealing with naturally occurring data and thus have experience in handling the associated concerns (Levallois et al., 2013), the new forms of data and the related research questions often elevate the associated problems to a new level.

First of all, in social data science the data often serve as a mere proxy for underlying phenomena of interest that are not observable (at least not practically). For example, when using social media data to monitor public reactions, one has to come up with operationalisations for what a reaction is and how a sentiment of that reaction should be interpreted. Lee et al. (2016) shows that, depending on psychological characteristics of an individual, pressing a "like" button on Facebook may have two distinct meanings – one is showing the appreciation of the posted content and the other is attempting to get affirmation from others. Moreover, even the forms of behaviour that do have direct analogues outside the virtual forms of interaction (such as dialogue) are also shown to be carried out differently online (Bruns and Stieglitz, 2014).

A closely related issue is that of sample self-selection: for social media, the data come only from the platform users; even for the most ubiquitous mobile devices such as smartphones, the data come only from those device users who opt into data collection. In both examples, non-participation is likely to be associated with a certain demographic profile (Ruths and Pfeffer, 2014), degree of technological savviness (Costello et al., 2016) and possibly other personal traits. Thus, any findings derived from naturally occurring data have to treated with caution when generalised outside the data-producing environment.

Natural occurring data may also be biased by a researcher's attempt to select relevant records. This problem will be discussed later in more detail within one of the case studies (see Section 3.3.4). In brief, it is a common task to identify the needles in a haystack of data, for example, posts on a particular topic among all social media posts from the platform studied. Two entangled problems may arise. First, the concept of a "relevant entry" is a fundamentally problematic one. For example, social media data are often conversational, so it may often be desirable to preserve the entirety of such conversations (Pennock, 2013; Voss et al., 2016). However, it may happen that only distinct posts within the conversation are on topic. Second, it is not trivial to operationalise the notion of relevance as a set of practical data selection criteria. For example, selecting social media data by hashtags has its limitations. Experience from my prior work

shows that some hashtags can be used in multiple contexts (Procter et al., 2015). Simultaneously, González-Bailón et al. (2014) show that accidental omission of some less popular topical hashtags may significantly skew descriptive statistics of the resulting sample of tweets.

Finally, even if some naturally occurring data are justifiably operationalised to the research questions and are acquired in accordance with appropriate data selection criteria, they still may pose research problem due to their limited quality. For example, there are severe concerns about the veracity of health data collected by the Apple smartphones: their built-in pedometers, while performing quite accurately in laboratory conditions, are shown to have an error margin of more than 20% in the wild (Duncan et al., 2018). In other cases, data may be accurate but contain plenty of missing values. For example, Twitter is known for having geolocation coordinates in only 5% of the platform's posts (Graham et al., 2014).

2.3 Ethics in Social Data Science

While explicitly mentioned only in brief, the theme of ethics in social data science has been underpinning much of the discussion above, as the notion of ethics is in general deeply intertwined with those of responsibility (cf. Keenan, 1997; Tauber, 2005; Knights and O'Leary, 2006) and politics (cf. Baron, 2001; Bellamy, 2010; Green, 2018) – not to mention the crucial role of ethics in modern research traditions. Many ethical issues in social data science have thus been already mentioned – and a more detailed review of them may be a subject of its own thesis. Yet, it is still worth to re-iterate that the public discourse around ethics in social data science and the corresponding ethical guidance and legislation are all in active development, most actively in the context of privacy. The famous suggestion by boyd and Crawford (2012) that "[j]ust because [data are] accessible does not make [them] ethical" (p. 671) immediately poses the question of what *does* make them so. As Voss et al. (2016) (citing Cate and Mayer-Schönberger, 2013) put it, there is a "question [of] whether or not ticking a box when signing up [...] constitutes an acceptable indication of consent" (p. 169). Even though some outline solutions – e.g. the one of social contracts (Martin, 2016) – are being provided, those are hardly definitive.

The changing landscape of actors actively involved in research (see Section 2.1) provides an additional layer of complexity. The scandal around the study by Kramer et al. (2014) that involved a mass experiment run by Facebook to manipulate emotional state of the platform's users showed some of the existing mismatches in assumptions about ethics – mismatch between the ethical expectations towards industry (and marketing research) versus towards academia (and non-for-profit research), and in data scientists' versus the public's understanding of consent (boyd, 2014; Felten, 2014). The ethical pressures on the industry are arguably "catching up"

with those on academia in the face of appearing new scandals such as the recent Cambridge Analytica one (Cadwalladr, 2017). The fast pace with which the discussion – and the sheer scope of the actors involved (Crawford, 2014) – unfolds can be illustrated by the fact that even academics (such as both boyd and Crawford in their respective pieces) often prefer to weigh in on the conversation in blog posts and opinion pieces rather in peer-reviewed publications.

The evolving nature of discourse on ethics in social data science has ramifications for the field well beyond moral ones – it affects the way the field is governed, for example through influencing the legislation. Arguably the biggest recent change in this regard is the General Data Protection Regulation (GDPR) that significantly increased the rights of European residents and, at least in principle, gives them control over the data organisations store on them (European Parliament, 2016). Such changes in legislation impact social data science both directly and indirectly. Directly, they may affect the logistics of handling the data collected for research. For example, GDPR forces researchers to track and rigorously record their decisions in regard to collecting, storing and archiving data (Dickson and Sigala, 2018). Indirectly, the data legislation changes cause usual external data sources to change their own data sharing policies and license agreements, which has arguably even more significant consequence for social data science than the direct legislation effect. For example, Twitter stopped returning user profile timezones via their public API in the face of GDPR (Piper, 2018).

In addition to formal legislation, there are also guidance sources that are not legally binding but may still have authority. An example of such guidance at the national level may be the Data Ethics Framework in the UK (Department for Digital, Culture, Media and Sport, 2018b). At the international level, the IEEE standard on algorithmic bias considerations (IEEE, 2017), which is currently under development, is a relevant example. Such guidance may influence the public discussion of ethics and may raise the bar of expectations towards ethics in social data science, thus driving more ethical research in both for-profit (through pressure from customers) and non-for-profit (through pressure from funders) sectors. This forms a system of *informal institutions* (Stiglitz, 2000) that govern the ethical landscape.

Overall, it is impossible to simply "deal with ethics and be done with it" in a social data science project. If research team members do not continuously ask themselves whether what they do is ethical – which also includes monitoring the landscape of the public discussion surrounding social data science and of the relevant legislation – they risk to accidentally dismiss some ethical dilemma simply because they fail to recognise its difficulty. This may not only negatively impact those affected by the research, it can also damage the researchers themselves, especially since they may be subjected to public backlash, to sanctions by their employers or even to legal consequences for their actions.

2.4 Research Guidance and Project Management

The discussion above shows that social data science is an activity performed in a highly problematic context of political, epistemological and ethical challenges. Those challenges are amplified by the logistics of social data science – project teams have to manage IT infrastructures, stakeholders engagement and their own work. Under such circumstances, keeping track of the research process and maintaining shared understanding within a research team is non-trivial. Yet, as shown in Section 2.1.3, such shared understanding is central to success in responsible social data science. This calls for provision of *guidance for social data science teams* that would help the teams to incorporate all the core aspects of a social data science project into their collective consideration.

Different pieces of guidance relevant to social data scientists are indeed appearing. Section 2.3 has mentioned examples in the realm of research ethics. Most of the appearing guidance focuses on methodology. For example, the book by Krishnan and Rogers (2015) guides businesses on modern types of social analytics. I have contributed to a handbook for social media research (Sloan and Quan-Haase, 2017). Finally, some limited guidance on management of social data science teams also exists. For example, Patil (2011) discusses the key competencies of social data scientists depending on their area of specialisation and suggests appropriate recruiting tactics.

Being focused on specific aspects of doing social data science, the examples of guidance above can be of great value when a social data science project faces a particular issue and the team member responsible for it requires additional advice. However, they are not tailored to observing the "bigger picture" of a project. Moreover, if the goal is not only to see this bigger picture but to share its vision across the whole team, a form of guidance that would spark *collaborative discussions* (Bokhour, 2006) would be optimal. Long policy documents and books are arguably better suited for a focused read than for interactive environments such as team meetings. A form of guidance that would be more appropriate to the tasks at hand is a **holistic project management tool**. Such a tool should provide a common talking point that all team members can use regardless of their disciplinary background. In that, it can help social data science teams to:

- consider the potential issues as early on in the project as possible;
- keep track of the project state, of the decisions that went into it and of their provenance;
- ensure that the whole team is aware of changes in project circumstances and of their effects on key project operationalisations.

While thus far social data science lacks its own systematic project management guidance that could be turned into such a tool, there are a number of disciplines that have provided substantive contributions to knowledge in project management and thus may potentially provide a starting point. For example, project management is (perhaps unsurprisingly) often a focus of management studies (e.g. Meredith and Mantel, 2012). More discipline-specific guidance appears, for example, in the field of industrial engineering (Kerzner, 2003). Arguably more directly relevant to social data science (and data-driven research in general) is project-level guidance from the field of research data management (Ball, 2012).

Software engineering (Sommerville, 2016) is another field that has been making substantive contributions to project management over the years, with these contributions being increasingly accepted across other fields (e.g. Richards, 2007). In the next section I will show that the fields of software engineering and social data science, despite their differences in the goals and nature of work, have a very significant overlap in key concerns of both hard (technical) and soft (logistics and management) nature. This overlap is crucial for successful adaptation of holistic guidance, as such guidance is meant to tackle multiple core project aspects at once. This makes project management guidance from software engineering *uniquely positioned* to be adapted to social data science. It is worth noting that this does not make project management experience of other fields irrelevant: the latter may be successfully adapted to tackle *specific* issues in social data science as and when appropriate.

2.4.1 Relevance of Software Engineering Project Management Guidance to Social Data Science

There are multiple reasons why project management guidance in software engineering tends to address concerns pertinent to social data science. To start with the obvious, doing data science in general often requires software development. Kim et al. (2016) argues that data scientist has become a legitimate role for a member of a software development team and points out that many data scientists write actual production code. However, even when software development is not required, social data science happens in the context of a project's software- and wider IT infrastructure (see Section 2.1.1). This is similar to how software engineering mostly "involves extension of preexisting systems and integration with 'legacy' infrastructure" (Finkelstein and Kramer, 2000, p. 6). Furthermore, software engineers do not necessarily seek to make such extensions through building software from scratch, as an organisation's needs may often be partially resolved through purchasing and integrating software (Daneshgar et al., 2013) and configuring the resulting system (Sommerville, 2008).

As much as social data science, software engineering requires careful stakeholder engagement

since the stakeholder needs inform the software requirements (Harker et al., 1993). The problem of understanding stakeholder needs and eliciting requirements has been long acknowledged and supported with extensive discussion (cf. Jirotka and Goguen, 1994). Active involvement of project stakeholders and recognition of their contributions is known to increase the chances of a software engineering project success (McManus, 2004) and recent attempts to involve wider social groups as project stakeholders (cf. Lohmann et al., 2009) only strengthen the parallels with social data science.

Arguably the most important similarity between social data science and software engineering is the non-linear process of work and iteration. As much as research questions, design and methods are iteratively refined in the face of limitations (limited availability of data, high costs of computation) and incoming evidence (interim findings), modern agile software engineering methods support continuous refinement of system design and requirements in response to the development iterations (Vlaanderen et al., 2011; Leffingwell and Widrig, 2000, pp. 225-310). These modern methods have become prominent in the face of the limitations that older approaches including the linear waterfall software development model (Aitken and Ilango, 2013). Hence, using them as a starting point to inform project management in social data science would allow to capitalise on the *maturity* of software engineering as a discipline.

2.4.1.1 Limitations of Software Engineering Guidance

The discussion above justifies applicability of software engineering project management guidance – and especially of the modern tools that adhere to the agile philosophy – to social data science. With that in mind, it is worth to discuss two characteristics common among the popular agile models that present a challenge for the process of their adaptation.

First, the models tend to suggest quite a rigid *way of working*. For example, the DSDM model (Stapleton, 1999) forces a defined set of roles on team members. SCRUM (Schwaber and Beedle, 2002) is more flexible and supports cross-functional development teams, however it implies a set of specific events through which the project is managed. Such strict methods may well be valuable for many social data science projects, especially the ones done in industrial environment under time pressure. More generally, however, the extreme variety of social data science projects calls for a project management tool that would be agnostic to such specificities.

Second, popular agile models tend to put less emphasis on the "soft" issues such as stakeholder engagement than required in social data science. For example, SCRUM suggests a single individual to bear the role of the product owner and effectively be responsible for delivering all the requirements to the development team (Schwaber and Sutherland, 2017). This is hardly possible in software engineering practice: in fact, Sverrisdottir et al. (2014) show that organisations use

different methods to resolve this issue depending on their individual circumstances. Arguably, for multi-stakeholder social data science projects having a "research artefact owner" is even less sufficient.

In response to the critique above, I suggest that a tool that could potentially be most successfully adapted to social data science should *provide a space of core concepts*[2] in which software engineering projects, their core aspects and the associated issues could be described – and should *not* try to construct a normative model for the way of running such projects. There is one tool that matches this description particularly well – the SEMAT Essence (OMG, 2015). Indeed, according to its underlying vision statement, the mission of SEMAT's efforts "is to identify and describe the elements that are essential to all software engineering efforts" (Jacobson et al., 2009, p.2). The following section will introduce this tool in detail and show why it is specifically appropriate to be adapted to holistic project management in social data science.

2.4.2 The SEMAT Essence

The Essence of Software Engineering (Jacobson et al., 2013) is a series of tools developed by SEMAT (Software Engineering Method and Theory) Inc. and accepted as an Object Management Group (OMG) standard (OMG, 2015). Underlying the SEMAT Essence is a *formal language* that can be used to describe the generic properties of projects in different fields. The language suggests to conceptualise project types through their *area of concerns* that contain core aspects (called *alphas* or *kernels*), suggest a set of activities to undertake (*activity spaces*) and require certain *competencies* to be dealt with. Figure 2.1 employs this language to represent the three areas of concern in software engineering, their core aspects (alphas) and their relationships.

The basic model of SEMAT Essence, which is also known as "Essence Lite" (SEMAT Inc., nd), breaks each of the alphas into a set of consecutive states through which the alpha has to go, so that the first state is achieved close to the project set-up and the last is (hopefully) achieved at the time of project completion. In turn, each state is assigned a list of conditions required for its achievement. This model allows to plan a software engineering project holistically by considering its many aspects from the start and continuously assess the progress of a project by tracking the states that different alphas have achieved. The model is particularly powerful at identifying "reverse salients" (cf. Hughes, 1993), i.e. relatively weaker progressed project aspects.

[2]The use of terminology here is inspired by the work of Gärdenfors (2000), who introduces the term "conceptual space". That said, I specifically restrain myself from using the exact term to avoid confusion with Gärdenfors's interpretation of conceptual spaces as geometric spaces in which mapping of concepts to their quality dimensions in human thought process can be represented.

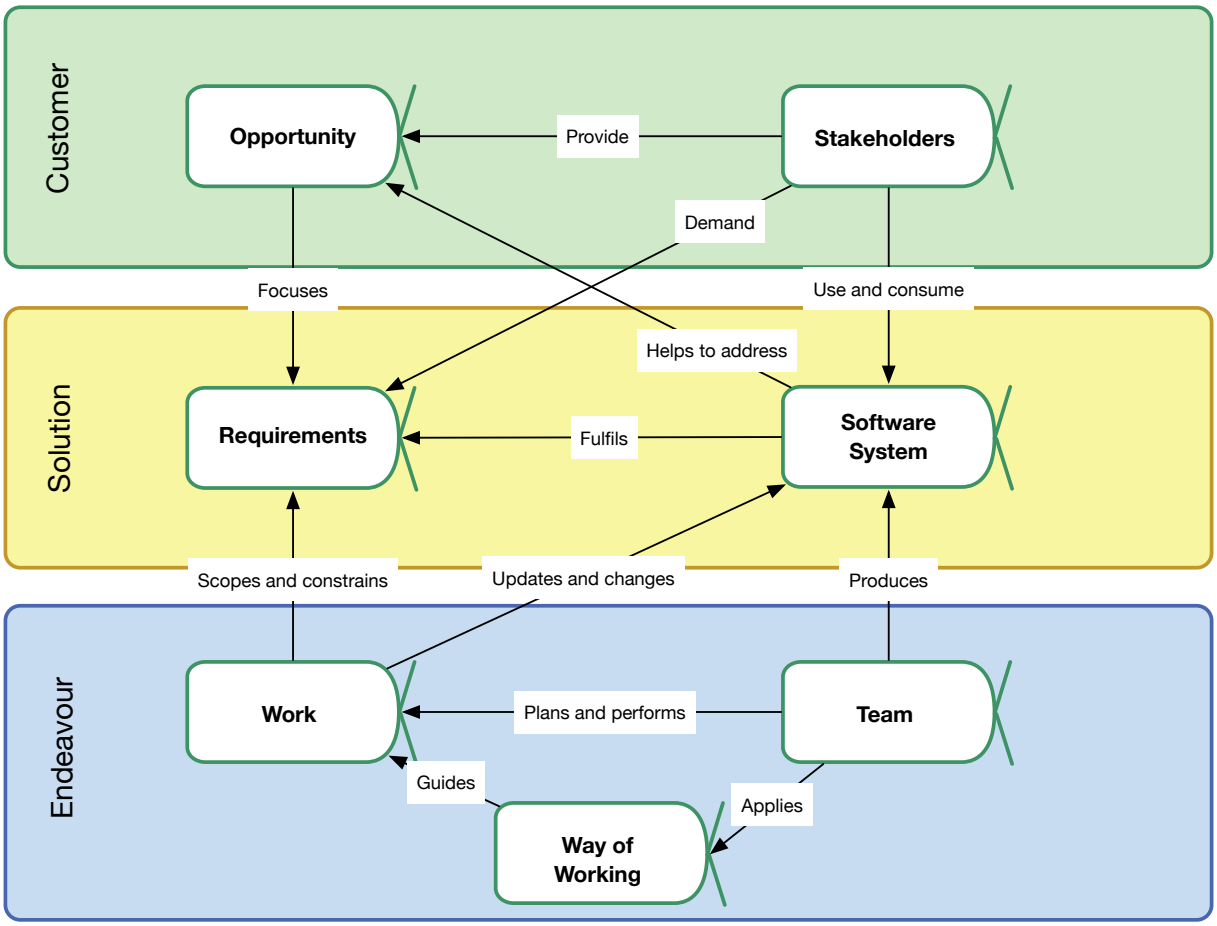


Figure 2.1: Essence of Software Engineering diagram. Adapted from the SEMAT Essence standard (OMG, 2015, p. 17). Provided under the Creative Commons Attribution License.

SEMAT implement this model as a set of cards for each alpha (listing their states) and each state (listing their conditions) that can be printed and filled in as an interactive exercise by a project team. Figure 2.2 provides examples of such cards.

The SEMAT Essence card deck is an excellent tool to adapt to social data science for two reasons. First, it is method-agnostic: by focusing on *what* has to be achieved, it omits the questions of *when* (temporal aspect) and *by whom* (division of labour). By virtue of including the "Way of Working" alpha, the Essence insists on adopting *some* method by a software engineering team – but what that method should be is project-dependent. Thus, the SEMAT model does not aim to replace method models – it is aimed to be used *in conjunction*. For example, there is an initiative for integrating the SEMAT Essence with SCRUM (Ivar Jacobson International, 2017).

Second, the SEMAT Essence does not simply acknowledge the existence of soft issues in software engineering project management, but rather puts them on par with the technical ones. The "Customer" area of concern contextualise the work on software development in the needs of

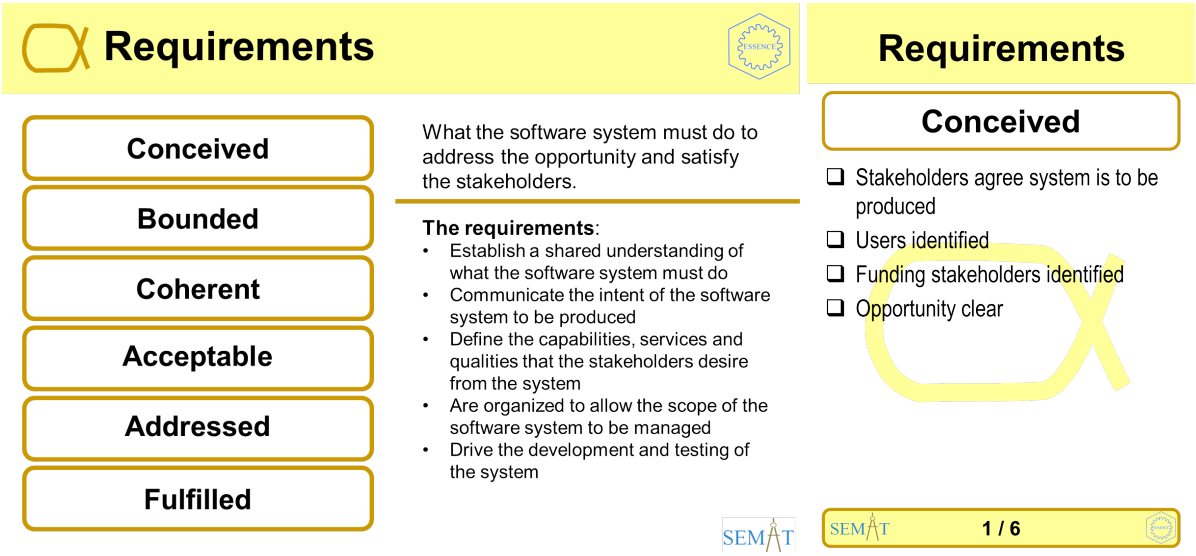


Figure 2.2: The "Requirements" alpha card and the "Requirements: Conceived" state card. Taken from SEMAT Inc. (nd). Provided under the Creative Commons Attribution License.

the stakeholders who require a software solution. The "Endeavour" area of concern looks at the logistics of turning the requirements into the software solution and tackles the issues of team management and organising the work. The technical aspects still get their representation through the "Solution" area of concern.

Overall, the SEMAT Essence model is generic, versatile, holistic and covers many areas of concern that intersect with those in social data science. It can be adapted to social data science by borrowing the overall structure (areas of concern – alphas – states – conditions) and, where appropriate, the content of the cards. I demonstrate the results of this in Chapter 4.

2.5 Chapter Summary

This literature review shows that social data science is an impactful activity performed in the context of social, political, epistemological and ethical challenges. Given this problematic context, social data science should be done responsibly. Research teams need to account for their decisions, as well as for the circumstances that led to them, and to maintain a shared understanding of those throughout the research process. Therefore, social data science teams would benefit from a project management toolkit. The SEMAT Essence model employed in software engineering projects (Jacobson et al., 2013) seems to be a particularly appropriate starting point to develop a project management tool for social data science, as this model covers many related areas of concern while not assuming any particular work process. The following chapter will inform the proposed project management tool through a range of case studies,

with each case representing a single social data science project studied through participant observation.

CHAPTER THREE

INFORMING THE PROJECT MANAGEMENT TOOL: CASE STUDIES

3.1 Methodology

Following on from the discussion of the need for a holistic project management tool for social data science, I now report on the fieldwork undertaken to inform the tool. One purpose of this is to discover previously unknown issues pertinent to social data science projects, to see how issues manifest themselves in practice, whether different issues are intertwined and if they amplify each other. The other goal is to study how these issues are dealt with within the parameters of real-life projects and their constraints – be that time, resources or experience.

The planned work is thus focused on *discovering* phenomena and describing them rather than on making testable claims. This strongly suggests using *qualitative, exploratory research methods*. The limited scope of the PhD studies implies that I had to make a choice on how much to rely on breadth-oriented methods that would allow gathering limited evidence from a broad range of social data science projects (for example, interviews or focus groups with panels of experts in social data science) and how much to rely on depth-oriented fieldwork methods that would allow me to observe a limited number of chosen projects. I made a choice to focus on the depth-oriented option and to employ case study design to inform the tool. The reasons are as follows:

- the goals of the work suggested a particular interest in the *longitudinal aspect* – not simply seeing the issues in social data science, but also how they evolve in time and how project

teams adapt to them;

- given the complexity and fundamental uncertainties (Grabe, 2005) of research, it seemed detrimental to rely on participant recall in the process of discovering issues in social data science projects – to employ an old metaphor, the devil was likely to be in details;

- orientation on breadth would most likely imply relying on the perspective of one particular project member in regard to every project, while in in-depth studies I could, through communicating with various project members, harness a more comprehensive picture especially valuable given the holistic nature of the suggested project management tool;

- ultimately, since I still had to make the final step of transferring the observations into recommendations embedded into the project management tool myself, I wanted to approach this step with *deep understanding* of the discovered issues that could be missed in breadth-oriented methods.

In this chapter, I describe several case studies, each representing a single social data science project. For the two primary case studies I provide complete ethnographies (see Sections 3.2 and 3.3), while two additional case studies (see Section 3.4) provide selected fieldwork evidence. While the details of how the full ethnographies were logistically carried out are provided in the respective sections, here I would like to start with providing an overarching background to the ethnographic work.

3.1.1 The Role of Prior Experience in Social Data Science

The ethnographic work I have undertaken as part of this research was supported by my past experience both in studying the individual disciplines that contribute to social data science and in participating in actual social data science projects. Being trained as an economist, I was exposed to almost extreme poles of the "quants and quals" spectrum of studying the social world. My undergraduate degree provided a strong background in empirical quantitative methods through modules such as Econometrics and Economic Statistics. On the other hand, I got introductory training in empirical sociology and was even exposed to social theory that preceded the prevalence of empirical method through a full module on "Das Kapital" (Marx, 1867). My MSc studies at the School of Computer Science enhanced my technical skills and provided me with basic training in machine learning (which I subsequently expanded through self-study), while the work on my MSc thesis found me doing empirical research through surveying and interviewing experts in the field of data journalism.

Furthermore, in the gap year between my MSc and PhD studies I provided consultancy to a social data science project where I performed social network analysis of Twitter data around the 2012 London Olympics (Willis et al., 2015), studying relationships among the participating actors and between actors on one side and the web content they shared as hyperlinks on Twitter on the other side. Another short research project reported in Hutchings et al. (2015) exposed me to doing mixed-methods research using Twitter and Facebook data around the Sochi Olympics in 2014. Indeed, this work had "interrelated components that might separately fit traditional qualitative or quantitative orientations" (Tashakkori et al., 2015, p. 620), as the quantitative analysis of the data was employed to identify the most influential postings and the most salient topics that in turn were then analysed through in-depth reading.

Thanks to this prior experience, I could engage in "immersive field work combining observation with participation" (Dourish, 2006, p. 542) from day one in each case study. In correspondence with the ethos of ethnography, I tried to not merely collect data on what I observing, but to engage in *continuous analytical interpretation* with an aim of *deep understanding* of the observed processes and with an end-goal of expressing that in *analytical reportage* (Anderson, 1994). The initial experience provided me with a head-start in understanding contexts of the studied projects. Additionally, I could be involved as a *participant* (i.e. a project team member) and study the work through first-hand experience. Finally, sharing at least some of the professional language with the rest of the teams facilitated engagement in meaningful conversations, while being predominantly quantitatively trained helped to quickly spot the language mismatches that I had with the qualitative analysts in the teams.

3.1.2 Ethnography for Design: Justification

As Anderson (1994, p. 156) points out, ethnographies are often "under attack by fundamentalism from left and right". Some attackers claim a *lack of generalisability* by denying that the depth and analytical nature of an ethnography can actually compensate for its inevitably limited breadth. Other attacks hold *intervention and subjectivity* against ethnographies: according to them, an ethnographer "distorts the data" on the observed social processes first through potentially influencing them through sheer co-presence (*ibid.*) – let alone participation – and then through providing a necessarily subjective account of what was observed (Dourish, 2006). Moreover, the defenders of ethnographic approach (cf. *ibid*; Anderson, 1994; Voss et al., 2009) question the premise of eliciting recommendations from ethnography. Below I provide a response to those critiques within the scope of this thesis and its goals.

3.1.2.1 Generalising from Case Studies

The issue of generalisation is relevant to my work. While the ethnographic case studies aim to stand on their own – i.e. to be valuable analytical reportage on the experience of social data science projects – they are ultimately used to inform a project management tool applicable in a variety of social data science projects.

First, the presented research does not claim the findings of the ethnographies themselves to be fully generalisable or "complete". The problem of induction suggests that empirical research is inherently incremental and relies on the replication of results by independent people and means (Gigerenzer, 2004; Vickers, 2014). I acknowledge the limitations of this study and and will be more than happy to see future efforts that could revise and expand its findings. What truly matters though is that this study *does provide new insights* into what it takes to do social data science.

Second, since the case studies serve a particular purpose of informing the project management tool, it is more important whether this *resulting tool* has wider applicability and usefulness. I attempt to evaluate these later in the thesis (see Chapter 5). The combination of the employed evaluation methods allow for both breadth and depth of the assessment, as the tool is appraised by a range of interviewed experts in social data science of different levels of seniority and is also employed for project management in an additional evaluation case study.

Yet, it is still worth discussing why the generalisability concerns did not prevent choosing the ethnographic approach to inform the project management tool in the first place. First, given the setup of this research, the tool can draw not on one project, but on a *range of case studies*. The case studies represent sufficiently different facets of social data science and thus "take account of probable relevant heterogeneity within the population" (Gomm et al., 2009, pp. 98-115) of projects, while retaining some of the key complexity dimensions associated with social data science – interdisciplinarity, variety of employed data sources, methods and infrastructural components, and management of stakeholder interests. Moreover, participant observation facilitates a very deep level of immersion in the studied projects that suggests a possibility to compare and contrast the findings from different case studies and determine which phenomena are caused by particularities of a study and which by the social data science core of the studied projects. In other words, it is possible to make *analytic generalisations* (Yin, 2012, p. 18). Finally, the case studies do not inform the project management tool "in vacuum" but rather in the context of literature on social data science (see Chapter 2).

3.1.2.2 Intervention and Subjectivity

Anderson (1994) groups the concerns of subjectivity and intervention together, since they both deal with the "purity of the data". As Dourish (2006) points out:

> "Ethnographic data are not unproblematically extracted from a setting, but generated through an encounter between that setting and the ethnographer." (p. 544).

Yet, a researcher plays a role in shaping the data regardless of the form of social inquiry. For example, when conducting closed-question surveys, a researcher influences the outcomes both through wordings of the questions and simply through the list of questions asked (Dillman et al., 2014) – and thus by their conceptualisation of the studied phenomena. Within the ethnographic approach, the influence of a researcher is not left "implicit in the data" (Dourish, 2006, p. 544), but is stated openly and critically reflected upon. Thus, when discussing the case studies, I provide detailed accounts of my relationship with the project and my role in the team (for example, see Section 3.1), indicate my influence on particular episodes (for example, see Section 3.3.4.1) and show how my perspective may be different from that of the other team members (for example, see Section 3.2.7.3).

Moreover, my involvement in the projects as part team member and part PhD student doing fieldwork allowed me to see both the bigger picture and the specifics of being involved in the projects. I would argue that both components are essential for deep understating. Judging from my past experience, without the general ethnographic aims in mind I would likely be preoccupied with my immediate tasks too much to make sense of the projects as a whole. Conversely, without being a member of the project teams, I would not have so much naturally occurring interactions with the other team members and would arguably struggle to internalise the motifs behind the observed work and the way of thinking that someone doing social data science may have.

3.1.2.3 The Issue of Is and Ought

Employing ethnography to formulate advice that should be captured in a project management tool is problematic in similar ways to using ethnography to gather requirements for software systems. Anderson (1994) and Dourish (2006), who discuss the value of analytical ethnography (as opposed to "scenic fieldwork" (Button, 2000, p. 319) in system design, point out that this value often does not lie in requirements specification and that "the most effective outcome of a study might be to recommend what should *not* be built rather that to recommend what should" (Dourish, 2006, p. 545). Voss et al. (2009) pose the problem in terms of "is" and "ought" – while

an ethnography provides a perspective on what the experience of participating in the studied setting *is*, by itself it cannot provide recommendations for what *ought* to be happening.

I do not seek to find a resolution to this conundrum. Rather, I accept the stance that a normative prescription has to come from an *argument* rather than direct observation. Still, there are two reasons why moving from "is" to "ought" in this research project is less problematic than in many others. First, the studied projects involved working in iterations. For some, the length of an iteration could be several months passing between the rounds of reporting (see Sections 3.2 and 3.4.1), while for one project the iteration length was one week that passed between two subsequent episodes of a data-driven radio show that the project was informing (see Section 3.3). The iterative work allowed to refine the process and resolve challenges discovered earlier. Therefore, the fieldwork provided direct evidence on how taking care of a particular project aspect could improve the overall research process.

Second, as the envisioned project management tool is to be based on the SEMAT Essence model Jacobson et al. (2013), the required "oughts" have to take a form of "what ought to be achieved" rather than "how it ought to be achieved". The lack of specificity on the "how" side means that, even if the way a specific challenge is handled in a particular studied project would not be not optimal in different sets of circumstances, the sheer facts of *handling* provides a strong idea of the respective "ought".

Now that the ethnographic approach to the fieldwork is methodologically justified, I can turn to the discussion of the case studies.

3.2 Evaluation of the Shakespeare Lives Cultural Programme

This section studies a research project that aimed to evaluate a yearlong Shakespeare Lives cultural programme led by the British Council. The project was hosted by the Open University but many of the team members, including myself, were consultants from a variety of UK institutions. The team evaluated the Shakespeare Lives programme across several dimensions with an emphasis put on the assessment of the programme's impact and value across different social media platforms.

Shakespeare Lives was an international programme centred on the commemoration of William Shakespeare and his legacy in the context of the 400th anniversary of his death. The programme started in early January 2016 and lasted until the end of that year. It constituted of a variety of online and offline events and activities including world tours of renowned theatrical troops, nationwide school competitions and the production of modern takes on Shakespeare's heritage such as short YouTube clips and Instagram picture series.

The programme was led by the British Council with support offered by multiple cultural, funding, and marketing partners across the globe. The British Council is a UK charity operating worldwide, whose declared aims include (British Council, 2016b):

- To foster multicultural cooperation in the fields of science, arts, sports and education;
- To provide educational opportunities worldwide;
- To provide engagement with the UK's culture, society, education and arts;
- To contribute to the international knowledge of the English Language and of the UK.

While the British Council is not a state organisation, it receives substantive state funding and it "supports the UK's national interests and priorities" (British Council, 2016a; Nye, 2004). Specifically, the British Council aims to promote the UK as a *welcoming* country that values *diversity, mutuality, innovation* and *creativity*.

As the centennial of William Shakespeare's death reinforced public attention to his art, the Shakespeare Lives programme offered a one-off opportunity for the British Council to reach – and potentially influence – new audiences. To seize this opportunity, the British Council decided to put an unprecedented (by their standards) large emphasis on using social media to host, disseminate and promote the contents of the programme. This included both creating content specifically for social media and using social media to raise awareness of the programme's offline

events. This is an uncommon practice for the British Council, as the organisation most frequently uses social media in a routine, not event-driven manner, e.g. to maintain educational public pages (for example, see the "LearnEnglish – British Council" Facebook page[1]). The nature of the Shakespeare Lives programme asked for an external evaluation project. The next section provides an overview of the project and, through this, sets up the case study.

3.2.1 Case Overview: the Evaluation Project

The evaluation project was commissioned by the British Council, who needed to have an independent assessment of their performance for internal purposes as well as to report to the joint funders of the programme, the GREAT Britain Campaign – the UK Government's international promotional campaign first launched in 2012[2] (see Section 3.2.3.1 for more details on the relationship between the British Council and the project team).

The project was run by a team of investigators:

- *The principal investigator*, who initiated and designed the project and oversaw its overall execution. She was responsible for administering the project funding and ensuring that the service agreement between the project team and the British Council was fulfilled ethically and responsibly. She was actively involved in supervision of all the ongoing work on the project, contributed to the project reports and directly led and conducted much of the research work. As will be shown below, some of it was outside the scope of this case study.

- *The co-investigator* with a background in computer science, responsible for oversight of the computational aspects of the involved work.

- *The co-investigator* with a background in political science who oversaw the issues surrounding *soft power* (Nye, 2004) – an approach to international relations focusing on establishing a positive image of a country among the populations abroad.

Aiming at comprehensiveness, the evaluation project included several research strands each pursuing its own objective. As the work on the three strands was, to a high degree, independent, the strands provide a useful framework to give a brief introduction to the project. I will outline the aims of each strand and briefly introduce their core methods and research teams.

However, before looking at each of the strands in detail, it is worth noting that I was deeply involved in the ongoing research- and reporting work conducted for Strand 1 during the

[1] https://www.facebook.com/LearnEnglish.BritishCouncil/

[2] More on the GREAT campaign: http://www.greatbritaincampaign.com/#!/about

Shakespeare Lives cultural programme itself and shortly after it finished. This work fits the definition of social data science (see 1) most closely due to the use of novel forms of social data (primarily social media data) and the associated challenges. It is the subject of discussion within this case study.

There was much work done for the overall evaluation project as part of the other strands. Moreover, the project investigators carried a lot of work after the social media analysis team was adjourned and the cycles of ongoing reporting were finished. All of those are outside the scope of this case study and are only introduced inasmuch as is required to better understand the work on Strand 1. Therefore, this case study should not be considered a full account of the Shakespeare Lives evaluation project, but rather a rich and comprehensive account of one of its particular aspects[3].

3.2.1.1 Strand 1. Qualitative analysis of social media interactions across different languages

Strand 1 of the evaluation project looked for evidence of the programme's influence on public perceptions of the British Council, Shakespeare and, ultimately, the UK on social media. For each of the five studied languages of social media interaction – English, Spanish, Arabic, Russian, Mandarin – and for interactions with the visual content on Instagram, the strand team investigated whether the social media users recognised the values that the British Council promotes (mutuality, diversity, creativity, innovation and welcoming nature). The team also provided an in-depth assessment of the public's emotional response to the Shakespeare Lives programme and examined whether the social media users found the programme's content enjoyable, useful and of high quality.

The key research method that the strand's team employed was content analysis actualised primarily through human coding of social media data and qualitative interpretation of the findings derived from the coding exercise. On each of the three rounds of reporting, a researcher coded a sample of social media entries from Twitter, Sina Weibo[4] or Instagram via a coding framework. The framework had been developed before the start of the project by the project investigators and was subsequently shaped by the strand team at the start of the project. In addition, the researchers took ethnographic approach to studying two social media platforms – Facebook and VK.com[5]. They were observing the public- and group pages relevant to the

[3] More on the overall evaluation project: http://www.open.ac.uk/researchprojects/diasporas/cvp/shakespeare-lives-2016

[4] Sina Weibo is a Chinese microblogging platform that resembles Twitter and Facebook. http://weibo.com/login.php

[5] VK.com is a Russian social media platform that resembles Facebook in its design and functionality. https:

Shakespeare Lives programme and the discussions that were emerging there over time straight through the platforms' web interfaces.

Given the high number of its research outputs, Strand 1 required a significant investment of labour. Indeed, for each round of reporting, five reports for data in each studied language and one report for the visual content analysis were produced, and a number of overarching final reports were prepared later. Data in each language was analysed by a single researcher (hereinafter referred to as *language researchers*) with a separate researcher being responsible for the visual content analysis.

The work of the six researchers was coordinated by a *project manager*. The project manager was *not* one of the project investigators, but rather another team member who provided their services to the project on a consultancy basis. Within the course of the evaluation project, the role was occupied by two different researchers of seniority similar to that of the regular researchers. The project managers oversaw the data collection from the third-party social media monitoring platforms (see Section 3.2.6.1) and played the key role in development of data acquisition criteria. In this capacity, they were a point of liaison with the British Council concerning some of the operational research decisions. At the end of the project, the project manager was responsible for meta-analysis of the findings derived across different languages.

The eighth and final member of the strand team was the *technical coordinator*. I took this role on the team. Within this capacity, I liaised with the analysts (i.e. the language researchers and the visual content analyst) on quantitative analysis and interpretation of data. I was responsible for supplying the analysts with data visualisations and for facilitating large-scale data collection when this needed to go beyond the social media monitoring platforms. The work of the strand team was supervised the project's investigators, who played the lead role in developing the methodology.

As mentioned above, due to the nature of Strand 1 and my personal involvement in it, the vast majority of the discussion on the Shakespeare Lives case study will be dedicated to the work on this particular strand.

3.2.1.2 Strand 2. Cultural Value of Shakespeare Lives: learning, monitoring and evaluation

The second strand of the research project aimed to provide an overarching evaluation of the Shakespeare Lives programme using the Cultural Value Framework previously co-developed by the Open University, the BBC World Services and the British Council (Gillespie et al., 2014).

//vk.com/

While the framework was adapted to the project needs, it retained its essence of assessing a soft power endeavour from the perspective of four groups of stakeholders: funders of the endeavour, collaborators, delivery teams within the organisation and the audience.

This research strand collated data collected with a wide variety of research methods and from different sources. The strand heavily relied on secondary data. For example, the "audience" component was predominantly informed by the work on Strand 1. Some of the data came from the British Council's internal reports and evaluations. However, this strand also generated new primary data – for example, through surveying British Council employees. The bulk of the research for this strand was done by the principal investigator and a senior consultant who had both previously led the development of the Cultural Value Framework.

3.2.1.3 Strand 3. Visualising Shakespeare Lives: interactive overview of the programme

The last strand of the project was concerned with creating visualisations to provide a high-level overview of the social media data that the Shakespeare Lives programme had generated. These visualisations were mostly aimed to enhance the presentation of the project deliverables and to support the qualitative findings of Strand 1 with relevant figures derived from larger-scale quantitative analysis. While playing an important role of illustrative support for the findings of the other project strands, these visualisations did not generate major new findings by themselves.

The work on the visualisations was done by a contracted third-party commercial organisation whose co-founder closely liaised with the project's research team. Additionally, sentiment analysis – i.e. automatic detection of the strength of tone underlying a piece of text (low to high) and of its direction (i.e. positive or negative) as expressed in a piece of text – of a large corpus of tweets containing the programme's *#ShakespeareLives* hashtag was carried out by one of the project investigators who used a readily available sentiment measurement algorithm, SentiStrength (Thelwall, 2017). Alongside the project's technical coordinator (myself), the investigator also lent support to preparation of the visualisations at the later stage of project. Example visualisations included maps and timeline diagrams depicting the spatial and temporal distributions of the relevant tweets and of their sentiment, an info-graphic demonstrating the most actively tweeting users and the most commonly mentioned Twitter accounts and an interactive calendar of the programme.

3.2.2 Fieldwork Methods

In line with the overarching methodology of the undertaken fieldwork (see Section 3.1), I had an opportunity to study the work on the project, and specifically on Strand 1, as a participant

observer. The study took the form of a range of activities in which I witnessed the work on the project in its naturally occurring settings. Table 3.1 summarises those activities across several dimensions:

- whether they were carried in physical co-presence with the project team members or distantly as an online ethnography;
- whether they included an active participatory element from my side or were purely observational;
- whether they coincided with my direct responsibilities as a member of the project team or were undertaken specifically for the PhD study.

Activity	Held in co-presence or online?	Included active participation?	Coincided with own responsibilities as a team member?
Project team meetings	Co-presence	Yes	Yes
Social media data collection and analysis sessions	Co-presence	Yes	No
Remote collaboration with the project team	Online	Yes	Yes
Study of project documentation	Online	No	No
Reflection on own work	N/A	Yes	Yes

Table 3.1: Fieldwork activities undertaken to study the Shakespeare Lives evaluation project.

The first type of activity was *participation in the project team meetings*. In total, I was invited to, and participated in four meetings. Two meetings with the Strand 1 team members were held at the start of the project. The first one was mostly dedicated to discussing the key features of the qualitative data coding frameworks. Those frameworks were subsequently used as the main analysis tools for Strand 1. I used this meeting as an opportunity to disclose my intention to use this project as a case study for my thesis and to ask the team for their consent. During the second meeting the team had an in-depth strategic discussion of the work ahead and pilot-tested the coding framework. The third and the fourth meetings were held after the first and the second rounds of reporting respectively. They were used to reflect on the work conducted for the rounds and, if necessary, to make methodological adjustments and revise project management practices. The fourth meeting also included the team members who worked on the Strands 2 and 3 to make a start on merging the strands' output into a coherent set of final deliverables. The meetings were neither audio- nor video-recorded in order to maintain their natural flow and, crucially, to not

disrupt their productivity. Written notes of the meetings were taken both by me and by the other project team members, all of which were re-examined for this write-up.

I organised three *one-on-one observational sessions* with the Strand 1 team members to get a detailed account of their work and of the issues arising. Two sessions were held with the project managers (as already mentioned, this role was held by two different people at different stages of the project). Those sessions were dedicated to selection and acquisition of Twitter and Instagram data. One session was held with the English language researcher. It was focused on coding the data using the coding framework. During the sessions, I encouraged the team members to think aloud and asked them questions to clarify aspects of their work. These sessions also involved active participation on my side, as the observed team members and I discussed the arising issues and resolved them together. I took written notes throughout the sessions. The data collection sessions were also audio-recorded with the consent of the observed project managers.

While the project meetings and the observational sessions required physical co-presence of the team members and myself, much of work on the Strands 1 and 3 took the form of online *remote collaboration* such as email exchanges and Skype calls, which I could observe only if I was participating in them or at least was copied-in. However, for a significant chunk of communication the team used an online project management platform Teamwork[6]. This platform supports sharing files and discussing topics among members of a project team. For this study, Teamwork was an invaluable resource. It did not only allow me to revisit the conversations that I had taken part in, but also to retrospectively examine the digital traces of collaboration among the other team members working on the Strands 1 and 3.

Finally, in addition to the highly collaborative forms of ethnographic work discussed above, it is worth mentioning two other methods used to study this project. First, as I had responsibilities as a member of the project team, some of the issues that arose in the project could be studied through *self-reflection* (Maréchal, 2010).

Additionally, to better contextualise the findings, *project documentation* was examined. This included the project proposal and the meeting notes taken by the other team members. The latter were especially valuable because they allowed getting an alternative perspective on the project meetings compared to my own. They also documented the moments of the meeting when my participation as a team member was more active and hence my ability to take notes was limited.

The fieldwork undertaken for this study yielded insights into numerous issues that were faced during the Shakespeare Lives evaluation project and into how the project team managed them (and with them). The discussion is split into subsections, many of which correspond with the

[6] https://www.teamwork.com/

alphas and areas of concern of the original SEMAT Essence model (OMG, 2015). The sheer fact that this grouping is logically consistent indirectly reaffirms applicability of the SEMAT approach to social data science – and even to such a relatively software-engineering-free project within it. Yet, the other issues relate to the project aspects not covered by SEMAT. Those, respectively, can suggest new alphas for the designed extension of the SEMAT project management model.

3.2.3 Findings 1. Research Goals and Stakeholder

The Shakespeare Lives programme evaluation project can serve as an excellent example of the importance to understand who the stakeholders are and to ensure that their goals are aligned with those of the team. I will show how different parameters such as a project setup and the differences in backgrounds of the project team's members and of the stakeholders may potentially make this non-trivial. I will also show how our team mitigated these obstacles.

3.2.3.1 Aligning stakeholder interests and team background

The first aspect to look at is the relationship between the British Council and the studied evaluation project. As is quite often the case in projects such as this one, the key stakeholder – the British Council – was simultaneously the subject of the independent evaluation carried by our project team and the body that had commissioned the evaluation. While our team was deriving the findings that (among other purposes) were meant to be used in reporting on the British Council's achievements to the organisation's funding partners, those further reports were the British Council's responsibility to prepare, and our team had no relationship (contractual or otherwise) with those funding bodies.

By the baseline guiding principles of independent evaluation, the British Council had no power over the findings that our team produced and could not reject our outputs if they satisfied the agreed level of rigour and quality. Being a publicly funded organisation with a reputation to maintain, the British Council was interested in outputs that complied with these principles. This was supported by our team for whom it was crucial to be compliant with good academic practice. The vast majority of the team members were either academics or doctoral students affiliated with various universities in the UK and abroad. Therefore, academic rigour was not only a matter of principle – it was also a matter of reputation and of the ability to subsequently use the findings to publish academically.

On the other hand, the British Council had had a long history of prior collaboration with the project investigators (e.g. during the development of the aforementioned Cultural Value Model); therefore, if the research outputs presented by our team had caused a conflict between the team

and the British Council, it could have jeopardised a long-established partnership. Additionally, the British Council could have *in principle* used some instruments of resistance such as coming up with formal reasons to not accept the work – even if, based on past experience, the team did not expect the British Council to do so. Indeed, the issue of pressure from funding bodies (including the well-minded ones) is not unknown in research. For example, Smith (2010) shows that UK researchers in health inequality who are funded to contribute to health policies feel bound to produce ideas compatible with existing policies. The following paragraphs will show how our team managed to maintain the integrity of the research within the project's setting – primarily through the efforts of the principal investigator.

The primary challenge faced by the principal investigator was in making sure that the research outputs were appropriately *recipient-designed*. Recipient design is a term coined by Sacks et al. (1974) for the process of formulating the message in a way that is sensitive to its intended audience. It is a recommended practice in qualitative social research as it helps to "avoid losing the audience" (Silverman, 2017). The purpose of recipient design is not to change the essence of the message, but to make sure that *the way the message is conveyed* fits the recipients' needs. For example, our team had to make sure that the findings were appropriately contextualised and that the overall tone of their discussion and presentation was tailored to the non-academic audience of the British Council representatives who might have been accustomed to less critical tone than the one often employed in academia.

As a result, the principal investigator often secured meetings with the British Council to demonstrate preliminary findings and to get their comments and suggestions as to what kinds of evidence our team might have overlooked in the research or what kind of questions could be answered in more detail. The meetings with the British Council were held in a close circle of people: the investigators and the project manager of Strand 1 from the project team side and the Shakespeare Lives programme manager from the British Council side. While this was not necessarily done specifically to reduce risks of confrontation between the British Council and the team, it achieved this as a side effect. The continuous feedback from the British Council allowed the project to stay relevant to its interests throughout. It was specifically helpful that the meetings with the British Council were consistently held in short advance of the project team meetings. Therefore, the principal investigator could communicate the British Council's relevant feedback to the team members in person and let the team thoroughly discuss it.

The principal investigator always maintained her strict academic integrity and was respectful to the aspirations of the project team. She never pushed to mask the reported findings – even the negative ones. Rather than that, she motivated the project team to ensure that those findings were appropriately framed, so if there were objective obstructions to more successful performance,

those would have to be discussed. Likewise, she made our team ensure that no evidence that viewed Shakespeare Lives in a positive light was missed. During the project meetings, she metaphorically encouraged the team members to "put two hats on", referring to the academic and the consultancy roles respectively. Given that the analysis mostly answered quite open questions with a support of predominantly qualitative research techniques, effectively the principal investigator achieved to motivate the team to dig deeper and come up with inventive yet appropriate research angles. As will be shown further when discussing the research methodology, this only strengthened the quality of the research (e.g. see Section 3.2.5.1).

3.2.3.2 Interpreting stakeholder needs

The long-term relationships between the project team investigators and the British Council were rooted in the investigators' interest and deep expertise in studying soft power organisations, i.e. the organisations that strive to establish a positive public image of a particular country among the populations abroad (Nye, 2004). Soft power organisations often strive to influence their audiences in indirect and tacit ways that ask for corresponding approaches to evaluation of their effectiveness. The British Council is a soft power organisation and the investigators' ability to theorise them as such was of extreme value for the Shakespeare Lives evaluation project. For example, it lay at the core of development of the evaluation framework for Shakespeare Lives, especially in the context of Strand 2. Therefore, both our team who assessed Shakespeare Lives as a soft power initiative and the British Council who acted as one attended to their *proper responsibilities*.

That said, in addition to having responsibilities (and thus interests) as a soft power organisation, the British Council had responsibilities as *just* an organisation. For example, the British Council had to report back to its funding bodies. The soft power nature of their activities sometimes made it difficult to pinpoint what their specific contributions were in such external reports. This occasionally led to potential difficulties of implementing some of the recommendations provided by our team. For example, during one of the data collection sessions designed to observe the team members at work, the project manager said the following about the tweets containing the British Council-initiated *#ShakespeareLives* hashtag:

> "There was stuff that wasn't necessarily directly related to the programme, but still used the hashtag; [the British Council would] be happy to see that, because for them that's still evidence of impact. All they want to see is that their programme shapes anything to do with Britain in any shape or form".

As a result, the final reports specifically praised the British Council's collaborations with local

cultural organisations, as well as the local events and activities that were done with very active contributions of the local partners. Further development of such activities was also one of the persistent recommendations across multiple project reports. While the British Council did not reject this recommendation, the organisation's feedback on our reports suggested that implementing this recommendation to the suggested degree could potentially create difficulties for external reporting. This example shows the importance of understanding the complexity and heterogeneity of interests of a single project stakeholder when reporting to them and providing them with recommendations.

Another point to make in regard to interpretation of the stakeholder needs is that the stakeholders themselves are heterogeneous. In the case of the studied project, while our points of contact were within the British Council headquarters in the UK, the list of the real beneficiaries of the evaluation project also included the local British Council teams across the world. There was one example of not accounting for the particular interests of such a local team (namely, the Russian team) in the initial stages of the project design, which was repaired in the early stages of conducting the research. This example directly links to the subject of data selection and thus is discussed elsewhere (see Section 3.2.4.1).

3.2.3.3 Expecting stakeholder engagement

Finally, to stress the importance of continuous engagement with the stakeholders, it is worth mentioning that lack of it once was a reason for miscommunication within the project team. The original project proposal promised a visualisation representing a "macro view on Shakespeare Lives in the format of a digital calendar" that should have been prepared by a third party contractor for Strand 3. There were no more stated requirements for this digital calendar, so the contractor expected collaboration with the British Council to elicit and refine those. However, the British Council focused its attention on the interim reporting done for Strand 1 and on the on-going work done for Strand 2, leaving Strand 3 out of the loop. This was most arguably due to the fact that Strands 1 and 2 actually produced output on a continuous basis, and the British Council's representatives felt more comfortable to engage in providing continuous feedback on interim results rather than actually participating in planning activities.

Simultaneously, some of the other work done for Strand 3 – specifically, the sentiment analysis performed by one of the project investigators – could also benefit from longitude visualisations, but in a form of tweet timeline diagrams rather than a digital calendar. The the focus of effort of the third party contractor naturally morphed into preparation of those visualisations with an implicit assumption that the timeline diagrams had replaced the calendar. It was quite a surprise for the contractor when, already after the internal deadline for delivering a working version of the

project website, they were contacted by the project principal investigator for the digital calendar. As it appeared, the plan to prepare the calendar always stood intact. The contractor did manage to deliver the calendar, but at high personal costs and on a hectic schedule.

3.2.4 Findings 2. Research Questions and Data Selection

Despite slight differences in interests and goals between the project team and the project funders (see Section 3.2.3), the overarching research questions of the project were clear from the start. These project questions, subject to minor variations in wording, were consistently re-iterated by the principal investigator during the project meetings. Specifically for Strand 1, from which I derive the materials for this section, those were:

- *RQ1:* What was the response to the Shakespeare Lives programme across the online platforms and languages? Was it positive, negative or neutral? Is there evidence that the social media users found the programme useful, enjoyable and of high quality?
- *RQ2:* Did the social media users acknowledge the values of Britain that the British Council aims to promote: diversity, mutuality, innovation, creativity and welcoming nature?
- *RQ3:* Is there evidence that the Shakespeare Lives programme affected the public's perceptions of Britain as a whole?
- *RQ4:* Which particular events and other efforts within the Shakespeare Lives programme were the most successful in terms of provoking positive feedback, promoting values and enhancing the international image of Britain?

While formulating the high-level research questions was straightforward given the goals of the evaluation project, it was quite a challenge to operationalise them to the level of a valid research design (De Vaus, 2001) – i.e., a clear understanding what data should be accessed and selected. In this evaluation project, it was not sufficient to clarify what data were required – we should have also considered what data existed in principle and what data were available for us. As a result, our data demands were refined iteratively throughout the three rounds of the Strand 1 research. This section focuses on the refinement of data selection criteria; the implications for data *analysis* are discussed further (see Section 3.2.5).

3.2.4.1 Selecting platforms to analyse

The first logical step in operationalising the research questions was determining what which platforms to use as the data sources. The original project proposal suggested analysing three

platforms – Twitter, Facebook and Instagram. This choice was primarily driven by the interests of the British Council and was shaped by the available resources. The British Council headquarters were strategically interested in Facebook, Instagram and Twitter analysis, as those platforms were most actively used to post on Shakespeare Lives in English. Two commercial social media monitoring tools that the British Council was in possession of and made available for our team also collectively provided access to specifically Twitter, Facebook and Instagram data.

During the first round of the project reporting, it became evident that the three chosen platforms, while being an adequate choice for analysis of the social media posts in English, Spanish and partially Arabic and Russian, were not appropriate for analysing the interactions in Mandarin. Neither Twitter nor Facebook were legally available to China's residents (Chiu et al., 2012). Although the first round of data acquisition was conducted right after the Shakespeare's death anniversary – i.e. after the highest spikes of the Shakespeare Lives-related online activity – we only identify less than a thousand relevant tweets to analyse. This number only became lower over time – e.g. only 271 relevant tweets were identified for the second reporting round. Likewise, there was not public/group page dedicated to the British Council in China on Facebook. For these reasons, a Chinese microblogging platform Sina Weibo (hereinafter Weibo) was added to the list of the analysed platforms.

Another social media platform that was added to analysis of the Russian language interactions was VK.com – the most popular social networking service in Russia. At the moment of this writing, it is the second most visited website in Russia according to Amazon Alexa rankings, while Facebook, with which is bears similarities in user interface and supported modes of user interaction, is only the 17th[7]. In contrast to Weibo, VK.com was included not due to necessity, but due to our commitment to improve research quality and usefulness where possible. Omitting VK.com from the analysis of Shakespeare Lives in Russian, while could be methodologically justified by the increased consistency of analyses across the studied languages, would have made the respective language report less useful for one of its key beneficiaries – the British Council Russia. The British Council Russia had established a very similar presence on Facebook and VK.com – on both platforms, they had a public page where they often posted synchronously. However, the VK.com group of the British Council Russia had twice as much followers as the Facebook group, thus the recommendations derived from the analysis of VK.com were of high relevance.

[7] https://www.alexa.com/topsites/countries/RU. Accessed on 17-09-2018.

3.2.4.2 Keyword-based and page-based data selection

While the choice of platforms to analyse required only a limited degree of refinement and was finalised during the first round of reporting, scoping the data selection criteria for each of those platforms was a continuous challenge throughout the project. For the language reports, the key factors that the project team had to consider included:

1. Affordances, user interface and structure of social media interaction on each platform;

2. Availability of data to acquire for each individual platform;

3. Specificities of social media interactions in each of the studied languages;

4. Specificities of the British Council's social media presence in each specific language.

Due to (1) and (2), it was decided early on in the research process that the scope of data selection would be qualitatively different for Facebook/VK.com as compared to Twitter/Weibo: the Twitter/Weibo data selection was focused on specific hashtags and keywords, while the Facebook/VK.com data selection was focused on specific public pages. This principle was maintained throughout the project's course.

Twitter is a microblogging platform with a restriction on the message length and without a hierarchical structure of pages, posts and comments. Twitter is known to raise questions on whether it should be considered a social networking service at all, as it is more similar to a short message broadcasting service (Kwak et al., 2010). Hence, it is appropriate (albeit with caveats) to treat tweets as individual postings rather than parts of a larger conversation. For this reason, a keyword-based data selection was deemed more appropriate by the research team. Indeed, such data selection can corrupt the structure of individual conversations (which is not as important for individual postings); however, with high level of confidence, it would return relevant data – or at least the postings that were deemed relevant by the users who used respective hashtags. A similar argument was made for Weibo due to its similar affordances.

By contrast, Facebook and VK.com are both social networking services that allow creating spaces for meaningful chained conversations. The pages-posts-comments hierarchy puts restrictions on visibility of each individual posting, so conversations cannot be hijacked that easily. More generous permitted post lengths allow for meaningful message exchange. For this reason, it seemed reasonable for the research team to focus the VK.com and Facebook data selection on conversations and other interactions happening on the public/group pages of various British Council country services as a whole, rather than on selecting individual message even if those likely be more relevant by themselves.

That being said, the user interfaces and the affordances of each platform were not the only factors behind the choice between the keyword-based and the page-based data acquisition strategy for each platform. The transparency and the privacy policies of the selected platforms also played their role. For example, Twitter is notably more open than either Facebook or VK.com. Both Facebook and VK.com support a range of privacy settings that limit who can see content and users do make use of this to limit the distribution of their postings, while Twitter has a simpler model and more people post their tweets publicly. While the social media monitoring tools provided to us by the British Council supported keyword-based data collection from Facebook, the resulting data samples would have suffered from self-selection, as only the posts within the unrestricted user profile spaces and the public pages would have been acquired. Also, the provided monitoring tools did not support VK.com data acquisition – and it was methodologically preferable to keep the data access approaches to VK.com and Facebook similar.

The considerations discussed above were a major factor behind the team's decision to take an ethnographic approach to studying Facebook and VK.com. Within this approach, the team did not systematically collect the data before the analysis – instead, the analysts observed how the conversations unfolded through the platforms' user interface and, if and when necessary, took screenshots as a form of evidence.

Weibo was a special case in terms of technical and legal availability of acquirable data. As might be expected from a Chinese social media platform, Weibo did not have an API that would allow performing programmatic systematic data acquisition. Moreover, commercial social media monitoring tools available to us also did not support any forms of Weibo analytics – let alone exports of Weibo data. As such, we had to rely on manual scraping of data returned by keyword search within the user interface of Weibo. Hence, we could not truly warrant completeness and/or representativeness of the resulting sets of Weibo posts. From this perspective, an account-based Weibo data acquisition could have been more robust. However, the British Council China did not have a Weibo account. This made identifying a reasonable set of Weibo accounts to scrape data from highly unlikely. Combined with the previously discussed methodological considerations, the keyword-based data acquisition from Weibo through keyword search and manual scraping was both a benefit and a necessity.

3.2.4.3 Formulating data selection criteria

The specific data selection criteria for Twitter and Weibo data significantly varied throughout the course of the project and, out of all the data selection parameters, deviated the most from the initial plan. Not only did the lists of keywords change, but the tactics and approaches to compiling the keyword lists differed from round to round of reporting, as much as from language

to language. The following will detail those differences with the focus on Twitter, as most of the Weibo data acquisition principles followed the Twitter ones.

The initial strategy was to focus each report on a list of specific most prominent events within the Shakespeare Lives programme that had happened before the respective reporting round and had provoked a high spike of social media activities in a studied language. We planned to look for unique tweets that were posted during those spikes and contained the *#ShakespeareLives* hashtag, as the British Council coined it specifically for the needs of the programme. We aimed to identify the spikes by examining the Twitter timeline diagrams produced by one of the available social media analytics tools. The ultimate goal was to select a random sample of 1000 tweets for each report so that the language researchers had sufficient yet manageable material for further qualitative analysis.

For the first round of reporting, we did manage to identify a sufficiently large spike of activity that happened across all of the studied languages – the aforementioned "Shakespeare week-end" in April, i.e. the week-end of the commemoration. Initially, we had a certain degree of confidence in the plan to use only tweets containing *#ShakespeareLives*. As the project manager said at the beginning of the first data acquisition session:

> "For [the British Council] internally, the key hashtag for anything surrounding this weekend was [...] *#ShakespeareLives*. And we knew, like, as soon as we started to look at it from the kind of observation perspective: [...] there is going to be a lot of traffic."

There was a lot of traffic indeed. However, it was only the English-tweeting audience who picked up the *#ShakespeareLives* hashtag widely enough that we could sample a thousand tweets. The Spanish language audience also used it to some degree, but they appeared to favour a different hashtag – *#Shakespeare400*. That hashtag was not directly linked to the Shakespeare Lives programme, but was also actively used and promoted by the British Council, so it was included into the Spanish data acquisition criteria with no hesitation.

For the languages whose writing is not based on the Latin alphabet, it quickly became clear that our team would either need to broaden the collection criteria in order to acquire the desired number of tweets or to moderate our appetites. It was primarily the project manager's decision to go for the first option. Prior to data collection, he contacted the Arabic language researcher with a broad set of English keywords for her to translate to Arabic. In his words:

> "It all depends on what we want in the end of this – do we want a very accurate

> dataset that might not reflect the full variety of discussion or do we want a very broad dataset that we can then sample from and get sense of broader conversation. I prefer the latter, which is why I probably gave [the Arabic language analyst] [...] as broad [set] of keywords as possible."

This decision of the project manager – and the fact that we as a team did not only agree with it, but also expanded upon it in the subsequent rounds of reporting – tells something about the implicit views of our team. Indeed, we could have argued for going with the easier route of adhering to the initial collection criteria and thus needing to analyse less data in the end. As we would collect and analyse all the tweets that satisfied the initial strict criteria, the small volume of tweets would not threat reliability of the analysis, since the acquired tweets would constitute the whole population-to-study (albeit a small one) rather than a sample.

Why then did we decide to create more work for ourselves? Reflecting of the past work, it seems that our team implicitly traced back to the initial broad research questions and thus treated the tweets we studied not as samples of a population of tweets satisfying the collection criteria, but as a sample of the much wider population of thoughts, opinions and and acts of self-expression of all the social media users who were affected by Shakespeare Lives. For such a wide population, its total size is principally indeterminable. Thus, we had a strong motivation to obtain a sufficiently sized sample. Interestingly, we never specifically discussed this within the team. I can express and rationalise it only as part of this post-factum analysis. However, the team conversations – including the project manager's quote above – clearly indicated that we all had this *tacit shared understanding*.

At the end of the day, the Twitter data collection criteria for Arabic, Mandarin and Russian were broadened with the word "Shakespeare" written both in English and in a respective translation. Also, specifically for Russian, one additional keyword ("Shakespereana") was added. This was the first time when an event-specific keyword was used as a data collection criterion: "Shakesperiana" was a series of school competitions in arts and languages. The competition prompted studying Shakespeare's legacy and were sponsored by the British Council.

During the subsequent rounds of reporting, we had to further increase the flexibility of the data selection criteria in order to retrieve a sufficient number of relevant tweets. While there were widely resonating events during the subsequent part of the Shakespeare Lives programme (e.g. Sir Ian McKellen's world tour as a Shakespearean actor), the resulting datasets were not sufficiently big for our goals. The new project manager, who stepped in for the second round of reporting, told me during a data collection session:

> "This time [...] for nearly all of the languages there is gonna have to be a lot more playing around because it's just there's far less obvious data available. [...] English is pretty much exactly the same as last time. [...] For every other language, there has to be some kind of cobbling together".

The *"cobbling together"* primarily took the form of acquiring multiple tweet datasets for each language. Each dataset was collected using keywords related to one specific event or a group of linked events. The smaller datasets were subsequently aggregated into bigger language datasets. However, each of the smaller datasets could also be used on its own for a more fine-grained analysis.

For Arabic and Mandarin, the project manager had to also include some deliberately broad data selection criteria in order to get a reasonable number of tweets. This resulted in a two-stage data selection practice in Mandarin. The Mandarin language researcher, before performing in-depth analysis of each tweet, had to go through all of them and select the ones that were, in her judgement, of relevance to the Shakespeare Lives programme. A similar procedure had to be applied to the Arabic dataset collected for the third round of reporting. Collection of that dataset coincided with the last and, arguably, the most drastic change to the data selection strategy. Due to the principle lack of Arabic tweets, the dataset contained tweets in English that were identified as coming from the Arabic-speaking countries by the black-boxed algorithms of the data collection tool. The Arabic researcher had to thoroughly scrutinise these tweets in order to make a conclusive judgement whether they were relevant to Shakespeare Lives and actually came from the respective region of the world.

For Facebook and VK.com, data selection was more consistent over time and was not prone so much to changing circumstances and to trial-and-error. Since a page-based approach to data selection was chosen for those platforms, the researchers examined the posts on the public/group pages of the local British Council teams. The only change happened on the last round of reporting. It was, once again, caused by the shrinkage in the amount of relevant social media interactions to study: Facebook/VK.com pages of the British Council's local cultural partners were included into the scope of the analysis.

3.2.5 Findings 3. Research Methods and Data Analysis

The Shakespeare Lives evaluation project provides a valuable example of adapting traditional social science methods to social media data analysis and thus bringing them into the realm of social data science. The bulk of work on Strand 1 was content analysis of the Twitter, Weibo and

Instagram data using pre-defined coding frameworks, while the study of Facebook and VK.com presented an example of online ethnography (Hine, 2015).

The project employed three different coding frameworks, each designed for a specific dataset type:

1. A coding framework for text-based social media data.

2. A coding framework for Instagram posts by the British Council.

3. A coding framework for Instagram posts by the wider audience.

Design of the coding frameworks was led by one of the project investigators with assistance from the principal investigator and the Instagram analyst. For the most part, unfortunately it was not observable for me. Each coding framework included a list of post attributes that a researcher should assess (e.g. tone of a social media post), their accepted levels (negative/positive/neutral tone) and a recommendation on whether this is a single-choice or a multiple-choice attribute (single-choice in the case of tone).

After the section on data selection, a reader may be surprised that the structure of the coding frameworks stayed consistent throughout the reporting rounds bar minor changes. This can be partially explained by the adaptation of previous analytic approaches from previous projects on soft power. However, another important factor was that the findings obtained through data coding could be enhanced through other, less structured modes of qualitative analysis such as deep reading. These findings could also be supported and clarified with example social media postings. Hence, anytime the coding framework by itself was not a sufficient tool to get a sufficiently comprehensive insight into the data, it was complemented by other methods. For this reason, maintaining consistency of this key analytical tool was given a higher priority.

Consequentially, all the fine-tuning of the originally designed structure of the coding frameworks happened during one of the two early project meetings and throughout the first round of data acquisition (i.e. before the first round of analysis began). The key issue was balancing the number of possible levels for each attribute, which is very similar to the bias-variance trade-off: few levels would mean low analytical precision, while too many would mean a difficulty to draw a line between different levels. Other issues were more of a technical nature, such as ensuring that appropriate "escape" levels ("NA", "Unclear", etc.) were provided and complex attributes were split into components.

That being said, the *application and interpretation* of the coding frameworks was not as set in stone as their structure. The next section takes an example of arguably the most ambiguous – and

definitely the most discussed within the project team – multiple-choice attribute in the coding frameworks: "values".

3.2.5.1 Understanding values

As mentioned above, one of the key goals of the evaluation project was to trace evidence (or lack thereof) of the British Council managing to promote the image of Britain in accordance with their five core internal values: innovation, creativity, mutuality, diversity and welcoming nature. Hence, it was almost mandatory to code for those values. As will be shown below, it is much easier said than done.

First of all, the British Council, while describing these values in their internal documentation, did not really operationalise them – or, at the very least, the organisation did not provide our team with operationalised definitions of those. Hence, from the very start, there was some confusion and ambiguity around the meaning of each value. For example, it was not exactly clear if the value of "innovation" was implied in a strictly business sense (i.e. sponsoring Research and Development, encouraging start-up activities, etc.). The team decided during the project meetings to interpret innovation broader, i.e. as proneness to using modern technologies. The "welcoming", "diversity" and "mutuality" values were not trivial to distinguish. While some conventions were developed – e.g. that "mutuality" required evidence of two-sided cultural exchange – the exact decisions still could vary from researcher to researcher and from context to context.

Another question was what to consider an appropriate acknowledgement of a value. Considering the British Council's mission, it might seem that only acknowledgement of a value in regard to the UK would be appropriate. Yet, no tweets with the *#ShakespeareLives* hashtag were explicitly praising the UK. This did not necessarily mean lack of *implicit* support. In fact, an explicit praise of the whole country in the context of discussing specifically Shakespeare would only seem wrong – and such a posting would more likely be sarcastic than genuine. The soft power influence is meant to be subtle and our team never aimed to limit coding for values only to direct acknowledgements of the values in regard to Britain.

If no explicit praise of the UK was to be expected, an alternative approach to detecting acknowledgement of the values had to be devised. One option was to require an explicit acknowledgement of the values in regard to the Shakespeare Lives campaign itself. We held a project meeting to discuss this after the first round Twitter data analysis – i.e. after each of the language researchers had a chance to code their first full sample of tweets. We decided that strong evidence of implicit mentioning also qualified. This notion allowed establishing shared understanding for some cases – e.g. that mentioning videos, theatrical performances and other

cultural outputs of Shakespeare Lives qualified as acknowledgement of the "creativity" value. However, for other cases it was still ambiguous what a *strong evidence of implicit mentioning* was. For example, it was not clear whether to consider a use of the Twitter emoji that portrayed Shakespeare an acknowledgement of British innovativeness. Since writing such a tweet involved using an innovative product designed by the British Council, the team decided to consider it valid evidence, but the questions was a subject to quite some debate.

The complexity further arose from the fact that the original list of the British Council's values was further expanded by a short list of values suggested by the principal investigator. While the British Council values were potentially attributable to Britain, the new values were potentially attributable to the Shakespeare Lives programme: the extra values asserted whether the analysed posts recognised the content of the Shakespeare Lives programme as "enjoyable", "useful/relevant", and "of quality".

As a result, our team had to develop its own practices to manage the risks associated with inconsistent understanding of coding for values across various researchers – and the risks of inconsistent understanding between us and the potential report readers. When writing up the reports, each researcher included a section that explained in detail how the values were interpreted and what was considered valid evidence of their acknowledgement. To minimise the chance of confusion, the reporters supported those discussions with representative example posts. By doing so, the team explicitly warned the potential report readers against comparing value-related statistics across the reports. However, since it was impossible to completely avoid the risk of such comparison, the researchers worked on establishing a shared understanding of how to code for values. The team dedicated much of the discussion time during the project meetings to the issue of value coding. Additionally, we held collective online discussions using the Teamwork platform where each researcher shared some examples of their value coding and the rest of the team gave them feedback. As will be shown further (see Section 3.2.7.2), the team had to limit their efforts in formal testing of inter-coder reliability – however, the evidence from the project discussions suggests that the level of shared understanding did increase over time.

3.2.5.2 Identifying "bots"

If ambiguity in coding for values posed a challenge for a particular aspect of analysing tweets and Weibo posts, the next discussed issue posed a question of whether some particular tweets should be considered for coding in the first place. In the very early stages of preparing for the second round of reporting, the project manager, while trying to collect the data on Shakespeare Lives in Arabic, noticed an interesting tendency – while some of the peaks in the timeline graphs of the Twitter activity corresponded with the events of the programme, others seemed to come out of

nowhere. Upon closer investigation into those random peaks, the project manager discovered a lot of tweets that did not seem to be created by people authentically interested in Shakespeare Lives. She had the following to say after all:

> "I ended up having to exclude some mildly pornographic keywords from the Arabic search. They were messing up everything. [...] Obviously because I don't speak Arabic at all it took me a long time to realise that so many of the tweets were spam-bot links."

This short example leads to a number of interesting observations. First, the methodological difficulties sometimes come from unexpected angles. I do not think anyone in our team could initially have predicted that spammers would be interested in hi-jacking such a relatively specialised topic as Shakespeare and commemoration of his death – yet they did. Second, critical engagement with data is valuable from the very start of the acquisition process. The project manager could not read the Arabic alphabet, so the timeline graphs were indeed the only way to look into the data she was equipped with. Still, this simplistic representation of data appeared revealing for her thanks to her critical thinking. Third, the team language sometimes develops tacitly. The use of the word *bot* in the quote above is not necessarily a correct one. Our team had no idea whether these hi-jacking posts were indeed a work of computerised bots or of human spammers. Somehow, the word "bots" just happened to be used to refer to the phenomenon – and it stuck for the rest of the project. It is an interesting question whether using the word "bot" instead of "spammer" when discussing specifically social media is a wider phenomenon. There seems to be a high interest to bots on social media in literature in the year of the Shakespeare Lives programmes (cf. Ferrara et al., 2016; Davis et al., 2016; Bessi and Ferrara, 2016), although this is hardly sufficient evidence to make strong claims.

While deliberate exclusion of tweets containing certain stop-words from the acquired data did help to reduce the numbers of the suspicious data points, it could not eliminate the problem completely. In some cases, the language researchers had to use their judgement informed primarily by indirect evidence. For example, during a data analysis session with a researcher of English-language discussions, she had to deal with the following tweet:

> **Tweet text:** *"William Shakespeare died 400 years ago today #StGeorgesDay #Shakespeare400 #ShakespeareLives | More: <a hyperlink>"*

The text of the tweet seems to be on topic. Yet, two factors raised suspicion. The first one was the "suspended" status of the users' account discovered through our attempts to access the user's

biographic information. The second factor was the hyperlink which proved to lead to some online chat service instead of anything relevant to Shakespeare. The language researcher and I deemed the account suspension and the promotional link to be sufficient evidence of the bot nature of this posting.

3.2.5.3 Interpreting findings

Even after solving (or at least tackling) the major methodological issues met by the data analysis, in order to produce the artefacts that would be useful for the British Council our team had to provide a meaningful interpretation to the findings. This proved to be a challenge of its own, since the terms in which the British Council, as a non-academic subject of evaluation, thought about the studied questions were quite different from the metrics that the "bare" analysis produced.

Indeed, our analysis provided quite direct answers to the questions as formulated in the beginning of 3.2.4. We had quantitative figures and qualitative characteristics for the online engagement with the programme, breakdowns of the success of individual events and so forth. Yet, what the British Council ultimately wanted to know was **"Is it good enough?"**. Answering such a question with academic rigour is rather more challenging, since at least some knowledge of what "good enough" constitutes is required. Unfortunately, for a one-off year-long international cultural programme organised by a partially state-funded soft power organisation there simply were no benchmarks. Even if there were some events that bore a degree of comparability, no open data were available for those – and again, there were not enough of those to formulate any kind of reliable statistics.

The way we tried to work through this challenge was by establishing the internal expectations of the British Council at the early stages of the programme. This was especially relevant for the Strand 2 of the research – establishing the Cultural Value of the programme – since the Cultural Value Model (CVM) is comparing different aspects of an evaluated programme with the baseline expectations. Yet, it quickly became evident that while the British Council had set targets for a limited number of parameters (arguably the ones their funders from the GREAT campaign had been most interested in, e.g. return on investment), most of the aspects assessed by the CVM had no targets. The attempts to motivate the British Council to produce those targets were also futile. Partially this of course could be explained by the British Council's unwillingness to put extra commitments on themselves. However, arguably even more so it was that they, as much as our team, had no idea how high to raise expectations, since Shakespeare Lives truly was a unique project. At the end of the day, this behaviour appeared to be the right choice for them, since some of the targets that were formulated were not met, clearly not from lack of effort but from lack of experience while setting them in the first place.

The question of "What is good enough?" was even more difficult due to the soft power nature of the British Council activity. Even if we could have formulated some targets for the impact of the Shakespeare Lives programme, we should have also moderated our expectations towards tangible evidence of achieving those targets. Most of the soft power's impact is by its nature non-tangible and cumulative. It does not necessarily consist of immediate reactions by the target audiences (such as approving social media posts) but rather of a gradual improvement or re-affirmation of attitudes towards Britain.

In the end, this challenge was tackled in two ways. First, as part of Strand 2 work, the notion of "goodness" of social media performance measured in Strand 1 was, if not fully operationalised, then at least broken down to its aspects, such as whether the social media users praised the programme as enjoyable, uniting, providing learning opportunities, etc. While these better-defined dimensions of goodness did not eliminate the fundamental underlying issue, they made making inter-language and inter-event comparisons easier. Those comparisons became the second part of the answer to the challenge. On the one hand, the analysts could have proposed a reasonably justifiable suggestions for what the British Council could have done better by (a) using the most successful parts of the programme as the baseline and (b) making necessary adjustments for varying scopes in the programme's particular events and in their audiences. On the other hand, the analysts could estimate how much tangible evidence of positive reactions to the different aspects of the programme events could have been expected in the best-case scenario within the programme.

It is worth noting that (b) was quite a challenge in itself. Not all the events were as discussion-provoking and could generate sufficient amount of evidence by themselves, hence the breadth of data collection criteria varied significantly from language to language and from event to event. At the end of the day, we had to assume that the analysts' degree of critical engagement with the data would allow them to mentally separate the effects of data collection from the underlying audience reactions in each case. Still, as a result, our team derived a competent estimation of the British Council's overall success with its social media audiences that incorporated multiple parameters with well-varied individual scores.

Interestingly, some of the analytical metrics derived from our analysis could be challenging to interpret by themselves. Moreover, if not carefully used or contextualised, they could even appear misleading. An example of that would be the sentiment/tone metrics. The distribution of sentiment across the social media posts was of high interest for the British Council, since this distribution *should* have, by intention, directly reflected public's attitudes towards the problem. For that reason, our team employed two methods of sentiment measurement. One was the above-mentioned automated sentiment analysis performed as part of the Strand 3. It was applied

to a wide sample of all collected tweets containing the *#ShakespeareLives* hashtag. While it was interesting as a first indicator, both our team and the British Council understood the limitations of the automated sentiment detection. Hence, the project investigators also put a sentiment-related question into the coding framework for human data annotation. The analysts had to code each studied tweet in terms of "tone" with only 3 broad value: positive/neutral/negative. Our implicit assumption was that combatting inaccuracies of automated sentiment detection with human annotation was the only required tweak to consider the sentiment analysis results at face value. As will be shown below, this assumption was not entirely correct.

The results of the tone analysis were, at the first glance, staggering: for many studied events, the vast majority of the postings were coded to have a "neutral" tone. This was very frustrating for the British Council. Such a finding could not tell much about the audience's reactions, and it also seemed a bit paradoxical, as the common sense suggests that people do tend to react either positively or negatively. However, a deeper dive into the data allowed to resolve this paradox. Most of studied postings simply shared information about the prospective events. Moreover, quite a lot of the studied tweets were generated automatically: some of the website that announced the programme's events had a "share this on Twitter" button, which generated an informative tweet with a link to the page and a short description of the event. For such tweets, the tone could not be anything else but neutral. Yet, while we understood that reposting information did not necessarily mean endorsement, mass sharing did indicate positive appreciation of the programme's events. From this perspective, marking those tweets as neutral did underestimate the public's sentiment towards the programme. In the later rounds of reporting, the overall tone figures were always accompanied by tone figures based only on the post that were not purely informative. This small example shows the importance of understanding that the metrics derived from social media data (and new forms of naturally occurring data in general) are usually merely proxies for the desired more tacit variables – and, as with every proxy, one should incorporate their limitations into the way they are interpreted.

3.2.6 Findings 4. Research Infrastructure

The work on the Shakespeare Lives evaluation project was facilitated by – and thus shaped by – a range of computer-based tools that formed quite a sophisticated technical infrastructure. While the infrastructure was continuously evolving to better suit for the changing demands of our work, the decisions on its key components were made before the start of the project. These decisions were as much based on the desire to fit the projects goals and questions, as on the resource constraints and available opportunities to gain elevated access to some of the tools. This section will mostly assess the role of those key infrastructure components for our work, but will also

tackle how we used other tools to tackle limitations we encountered.

3.2.6.1 Acquiring data with social media monitoring tools

Our team engaged in the evaluation project under an agreement with the British Council that the latter would provide us with access to the social media monitoring tools used by the British Council themselves – Sysomos and Brand24. Such tools are mostly used for the purposes of online marketing and brand monitoring, and they provide elevated access to social media data[8] and built-in analytic capabilities. The latter were of little interest to our team for two reasons. First, the analyses offered by those tools tend to be methodologically opaque with the documentation providing only limited suggestion for how to interpret the findings (Procter et al., 2015). Second, the British Council were most interested in forms of analysis that they could not run themselves. That being said, our team did need access to Sysomos and Brand24 to acquire the social media data. For example, with Twitter, the only real alternative would have been to collect the data via a publicly available Streaming API. That would have meant having at least a hardware setup that could be run with minimal interruptions 24/7 throughout the course of the project with continuous access to the Internet for real-time data collection and, perhaps more importantly, good *ex ante* understanding of the required data collection criteria. The former could be quite a resource stretch for us, while the latter was simply not possible since data acquisition was iterative and event-driven.

While these tools did allow us to access historic data on tweets and Instagram posts[9], we faced how little control we actually had over those tools. During the first data collection session, the project manager was dealing with a deluge of *#ShakespeareLives* tweets in English. To export these data from Sysomos, he had to break down the retrieval criteria into chunks, since Sysomos had a cap on the volume of tweets per one export. The fact of this cap's existence was not a surprise by itself – indeed, the manager had recently used Sysomos in a different project. What he did genuinely not expect though was how low that cap was – only 5000 tweets per single export versus 30000 at the time of the previous project. Interestingly, I also had had a prior experience of working with Sysomos at an even earlier moment. At that time, the limit stood at 15000 tweet per export – yet another value for the same variable. Sysomos was simply changing the level of service that they were providing without much choice for their customers. While such fluctuations in service did not cause hurdles specifically in the Shakespeare Lives project, the observation did raise a question of how much the commercial social media monitoring tools could be relied upon in a long-term research project.

[8] compared to publicly available APIs.

[9] For example, Sysomos has full Twitter data for one year.

The data selection functionality of both Sysomos and Brand24 was also severely limited. For instance, Sysomos did not allow to explicitly filter out retweets. That was highly desired, since our team aimed to analyse as broad a set of audience reactions as possible; retweets thus were a danger for the qualitative scope of the analysis. We did develop a workaround for this limitation by exploiting Twitter's internal representation of retweet texts (starting the retweet text with "RT @<user name>"), but this solution was also not optimal since we potentially could exclude original tweets that happened to start with "RT"[10]. Brand24's data selection capabilities were even weaker, as the tool did not support boolean search (although it did allow access to posts of a selected account). As the lack of boolean search functionality did not allow for formulation of sophisticated queries, it kept the team from iterative development of the data acquisition criteria on Instagram – we simply employed the *#ShakespeareLives* hashtag throughout the programme.

It is worth to make several remarks regarding the discussion above. First of all, a limited data selection functionality diminishes the value of social media monitoring tools even more significantly in their default use case: data analysis (not retrieval). Since our team exported data for external analysis, we could apply some more sophisticated filtering after exporting. By contrast, those who use the monitoring systems for data analysis have to accept that their analyses will be applied to sub-optimal selection of data points. Second, Sysomos currently claim to have a more powerful selection functionality: the system allegedly allows to query for specific topics rather than mere keywords (Sysomos, nd). While this may work to some extent, this even further reduces the users' control over what data are fed into the analysis. While one can argue that the users can eyeball the retrieved data and judge whether they are actually on topic they've queried, this does not really work at scale and this does not prevent false negatives. Besides, if the user did find the retrieved data irrelevant, what could they do besides trying a different topic name?

Some of the limitations of the data retrieved with the employed commercial tools actually motivated us to expand the acquisition infrastructure beyond the initial plans. For example, the Twitter data retrieved from Sysomos lacked users' self-stated short profile biographies. A user's bio was of primary importance, for example, to distinguish whether the tweet came from a member of the programme's audience or from one of the British Council's cultural partners. The analysts could have manually look up this information on Twitter – Sysomos data did include all the required links and the profile names – but having to go to Twitter for each new analysed tweet would have been inefficient. As a result, I acquired user biographies for all the tweets to be analysed via the Twitter REST API. A similar problem arose with the links shared on Twitter. In the internal representation of tweet texts, such links are shortened using the "t.co" domain name.

[10] We did do a quick analysis that showed that such tweets were practically non-existent.

Sysomos data contained only such shortened versions of links, so assessing them basing solely on Sysomos data was impossible. The analyst had to either follow the a link through, access the original tweet to see the link in full or discard the link.

Finally, at the time of our project, neither Sysomos not Brand24 provided appropriate functionality to access data from the studied local social media platforms. This was not a problem for VK.com, since it was studied ethnographically. However, for Weibo, the Mandarin analyst effectively had to rely on the platform's search engine and to manually scrape the relevant data. Since the volumes of discussions on Weibo were never overwhelming – and thus the researcher could exhaust the search and go through all the posts retrieved – we could be confident in the outcomes. Yet, it was a lot of tedious work.

3.2.6.2 The price of repurposing

As mentioned above, Sysomos is aimed at use in a commercial, marketing context to monitor a brand's success and to gather the aggregate sentiment of its audience. The use of this tool to export data for other modes of analyses performed at an academic level of rigour was essentially a case of repurposing – although a very mild one. There was not much for us to do to tailor the tool to our needs that would go beyond coping with the tool's limitations discussed above. However, there was one particular problem that was very relevant for us as academics – *tracking data provenance*, i.e. the decisions and the actions that led to each of the studied datasets.

It was crucial for us to preserve the criteria that fed into each of the analysed samples of social media posts as the data collection criteria varied drastically from sample to sample among both languages and rounds of reporting. Moreover, for the sake of research quality and confidence in our findings, it was equally important to preserve as least some metadata on all the rejected iterations of data retrieval criteria: at the end of the day, it was the iterative, trial and error process of appraising different datasets and adjusting the data retrieval criteria that justified what data the researchers had to analyse. As the first project manager put it:

> "One thing I'm very keen on doing with this project is maintaining [... a] paper trail of everything I do. [...] It includes date range, keywords searches. It also includes number of posts in the sample".

As expected, Sysomos did not support any functionality that would aid the project manager in tracking data provenance. Therefore, he had to develop a format for manually editable metadata records. He kept the *"paper trail"* in a Microsoft Word document, with each single metadata

record implemented as a small table of a fixed layout. Such tables had to be manually created and filled in by him each time he queried the data.

The resulting Microsoft Word documents indeed provided quite a comprehensive trace of the data collection activities. However, as is often the case with manual bookkeeping efforts, the key problem was having the discipline to record metadata at each and every search. In the first round of data collection and reporting, it was already tiresome – but not quite as much as in the subsequent rounds. Indeed, during the later rounds our team faced a lack of data retrievable by the more obvious data collection criteria. Hence, the acquisition process had to involve even more iterations. The project manager of the later project stages confessed:

> "It's very difficult to keep track of [laughs] everything that's going one: what I have done and what I haven't done."

3.2.6.3 Annotating data in spreadsheet software

In contrast to what a reader with a social science background might expect, our team did not use qualitative data analysis tools such as ATLAS.ti and NVivo for human coding of Twitter, Weibo and Instagram posts. Instead, three types of coding framework were prepared as stub Microsoft Excel spreadsheets – i.e. files with all the required columns and data validation, but no data – that were populated with data from each language / platform for each researcher. This may seem like a counter-intuitive choice, but it was in fact grounded in a number of good reasons.

1. *Reuse and experience.* At least one of the coding frameworks used in the project was heavily based on a framework developed by one of the investigators for an earlier project. That framework had been implemented in Excel, so sticking to Excel did not only mean an opportunity to reuse its stub without format conversion, but also to reuse the overall workflow of that past project.

2. *Perceived affordances.* Since data coding was meant to strictly follow a predefined framework, it made sense to create an environment for the researchers that would remind – and motivate – them to stick to the predefined codes. An Excel spreadsheet with a separate column for each variable-to-code and strict data validation applied to each value cell arguably was effective in creating the image of enforcement to "play by the rules" – even if in practices those rules could easily be bypassed.

3. *Coding granularity.* This point is related to the one above. Some of the key advantages of the qualitative data analysis tools lie in the opportunities to apply coding to extracts of texts on an arbitrary level of granularity – e.g. to words, sentences or whole paragraphs. However, in

the task at hand, coding was always applied to the complete social media posting, so tabular interface was perfectly suited.

4. *Format compatibility.* The exports from Sysomos and Brand24 were in CSV format. To effectively present these data for coding in the interfaces of ATLAS.ti or NVivo would have required representing them in a more readable – yet textual – form. Most likely that would have meant writing a script to populate HTML files with the data and formatting them with CSS. Using spreadsheet software eliminated this need.

5. *Ubiquity.* Even though many analysts on the team came from humanities or qualitative research oriented branches of social sciences – the disciplines that for the majority of the time do not require spreadsheet software – everyone to have access to some version of Microsoft Excel either through personal resources or through the resources of their respective institutions.

This example shows how tool selection can – and should – be effectively tailored to a project's needs and circumstances rather than to the default solution associated with a particular research technique. The use of spreadsheet software did prove to be the best fit for the task at hand, to the point that our principal investigator retained this practice in her future project that I took part in [11].

3.2.6.4 Automating data visualisation

Arguably the biggest cost of using Excel instead of specialised qualitative data analysis tools was the lack of automated reporting and visualisation functionality for the coded variables in Excel. Of course, it was possible to produce very neatly formatted diagrams in Excel and, furthermore, those diagrams would have covered almost all the chart types required to summarise the results of human coding. The problem was in producing those diagrams at scale. Indeed, the Twitter / Weibo data on each round of reporting were coded for 5 languages across 10 variables. Many of the variables were multiple choice, each choice thus being reflected by a single spreadsheet column. Considering the fact that some of the charts should have covered not only individual variables, but also their relationships, the total required number of diagrams could reach several hundreds for each round of reporting. Producing charts in such quantities using the Excel user interface was impractical.

Since this was understood early on in the project, it was agreed that as part of my role as a technical coordinator I would have to come up with a solution for effective plot generation. I

[11] Evaluation of the "InfoMigrants" information platform (see Section 3.4.1). We switched to Google Sheets from Microsoft Excel for collaborative editing and better support of Unicode.

developed a library for generating relevant plots in R using the popular *ggplot2* visualisation package. Developing that library appeared to not only be of logistical value, but also *research value*. Indeed, the library ended up influencing the research workflow and the role that plots played in the research.

Originally, the plots were supposed to merely illustrate the reports that the researchers prepared on the data they had been annotating. The analysts provided me with requests for the plot types that would support their narrative. Generating the plots that were formatted to publication standards was simplified thanks to the developed library but still not fully automated. However, the plots that were "quick-and-dirty"– yet still perfectly readable for someone who was familiar with the underlying data – could be generated fully automatically and at zero extra cost. This led to an idea of producing an exhaustive list of exploratory plots for each researcher. Those plots were to represent distributions for all the individual variables, as well as correlations within all the pairs of variables. The researchers could use such plots to generate further ideas for their reports – and only then request properly formatted plots when required. This workflow was especially successful since each researcher was to prepare a report specifically on the data they had coded and thus had great familiarity with. Thanks to this knowledge, if an exploratory plot flagged a spurious correlation rather than a substantive relationship, it was confidently dismissed.

3.2.7 Findings 5. Project Team and Collaboration

As is evident from the discussion above, the evaluation of Shakespeare Lives was grounded in data from different sources and in different languages that required a broad mix of methods to analyse. Variety of the research tasks at hand implied a broad set of skills and competencies required from team members. As a result, our team was interdisciplinary (with team members from humanities, social science and computer science) and international (with team members residing in the UK, Germany and China). Such a variance of members' backgrounds and circumstances is often thought to be problematic for team collaboration (cf. Klein, 1993; Henderson, 2005; Bracken and Oughton, 2006; Luthans and Doh, 2015).

On the other hand, the required level of team coordination was high. Even though each team member was highly autonomous in the tasks they were working on, they all depended on the tasks perfomed by each other. For example, while each researcher worked on reports for a specific language on their own, they all required input from the project manager and the technical coordinator in the form of data and visualisations, and they all had to cross-validate and calibrate their findings among each other so that their work could be effectively synthesised by the the project investigators, manager and technical coordinator.

Given all the above, it may seem to the reader that the team collaboration should have been difficult to say the least. Yet, against all odds, our team was consistently collaborating smoothly throughout the course of the project. In fact, the principal investigator and I even had an informal discussion after the end of the project on the possible underlying reasons of this ease. Some of those may be:

- *The efforts of the first project manager,* who, according to the principal investigator, "did a marvellous job" in setting the project up. The project manager's tasks were originally set as high-level aims (e.g. to make sure that the researchers have the required data and that they produce analyses in accordance with the project requirements). He turned those into a workflow and *de-facto ways of working* that served us through the subsequent rounds of reporting and developed the aforementioned provenance tracking system that supported this workflow.

- *Collaboration modes and tools,* which we had a big arsenal of. As mentioned earlier, in addition to the usual modes of remote collaboration such as Skype calls and emails, our team greatly benefited from a series of face-to-face project meetings and the communication within the Teamwork project management platform[12]. The role of each specific communication mode will be discussed in greater detail below.

- *Work ethic of the team members,* who all were committed to the idea of doing quality research. In the project meetings, it was not uncommon for the whole team to critically engage in challenging discussions. In the observed data collection sessions, both project managers went the extra mile to collate comprehensive yet relevant datasets. In the observed data analysis session, the analyst of English tweets took extensive notes about the most interesting tweets she encountered in addition to coding them. Most analysts at some point initiated new discussions on Teamwork to share their experiences and establish best practices.

The following sections will go through the most collaboration-intensive aspects of work – formulating data collection criteria, ensuring inter-coder consistency and selecting modes for quantitative data representations – and will look both at the most critical obstacles for effective team collaboration that we had to resolve and how we employed different modes of communication to do so.

3.2.7.1 Collaboration barriers to data collection

While most of the day-to-day activities in the project were performed autonomously, distance and time barriers sometimes were a big hurdle for collecting Twitter data. This was arguably

[12] https://www.teamwork.com/

the most iterative process in the whole project; ironically, it was also the one that required the closest collaboration between the project manager and each individual language research. Indeed, both project managers employed over the course of the project were English speakers with little to no knowledge in other languages for which they had to collect data. Therefore, the input from language researchers was required. While the project managers did their research into the Shakespeare Lives events happening in different countries and could use dictionaries to formulate respective data acquisition criteria, it was quickly established that without feedback from a respective language researcher those criteria were insufficient. First of all, the project manager could accidentally omit morphological forms and relevant cognates. Second, some of the relevant criteria were non-dictionary terms that appeared around the Shakespeare Lives events, for example the aforementioned "Shakesperiana" – the name of a Shakespeare-themed school competition in Russia. Additionally, the acquired data had to be appraised by at least eyeballing to judge whether they are indeed relevant. This was somethings automated translation could be of more help with, but an additional oversight by a language speaker was crucial.

In the absence of co-presence, online meetings via Skype were the best way to quickly perform the required iterative collaborative work. However, since every language researcher had other study and work commitments, the iterations over data collection criteria often had to go via asynchronous communication tools such as email or Teamwork. Because of this mode of communication, a bit of back-and-forth Google Translation was still required. For example, when sending the updated sets of data collection criteria, the analysts did not always have time to include the translations for them. It was often quicker for the project manager to double-check those without the analysts' assistance using automated translation. Furthermore, in a rare occasion of a Skype call between the project manager and an analyst *did*, the project manager tried to maximise the utility of the call and make it useful for more than just its immediate purpose. For example, in a conversation with the English analyst before the third round of reporting, the project manager tried to not only specify data selection criteria for English, but also to get some benchmarks for what kind of data to acquire for other languages.

Even in cases of email communication, project manager and language analysts often still agreed on the date when the relevant data were to be collected, so that the analyst would check their inbox messages more often and try to provide timely feedback. From this perspective, it was most problematic to iterate through the data retrieval criteria in Chinese, since the timezone difference with China often effectively meant getting feedback the next day. This arguably was an additional tacit reason for why the Chinese researcher mostly switched to Weibo data, which she collected herself, for coding on the later stages of the research. While she continued to analyse Twitter data as well, less emphasis on these data allowed her to be a bit less thorough in

deriving the best possible Twitter dataset, thus reducing the number of required iterations.

3.2.7.2 Collaboration barriers to inter-coder consistency

Another consequence of the language barrier was the difficulty of establishing inter-coder consistency. While everyone could appraise the efforts of the English-language analyst, the rest of the languages were opaque. The overall strategy that our team employed to combat this obstacle was to put more effort into establishing a consistent approach to data coding *ex-ante* and into maintaining it throughout the coding exercise rather than into the *ex-poste* assessment of each others' coding results.

The face-to-face project meeting were the key medium to implement this strategy. Since our team was geographically spread out and the number of the meetings had to be kept small, it was very important that each of the meetings was of high value for the team. In accordance with the selected strategy for assuring inter-coder consistency, in an early project meeting that took place before the first round of data coding, we all – not only the analysts, but also the project manager and myself as a technical coordinator – pilot-coded a handful of tweets in English, discussed the disparities in the outcomes and reached an agreement. This helped to establish a good level of shared understanding of the coding framework from the start of the project. This was effectively sufficient to establish a required level of inter-coder consistency for all categories bar the most challenging ones, such as the aforementioned "values" of the British Council. The next project meeting that took place after the first round of data coding was largely dedicated to reaching agreement specifically on those more problematic categories.

While the two project meetings discussed above allowed to establish the required level of shared understanding, the team still had to make sure that this understanding was maintained throughout the project and was put into practice. For this, the project discussion boards on the Teamwork platform became an invaluable resource. For each of the more problematic coding categories, the researchers shared their translations of the specific tweets that they either were not sure how to treat or they thought were the most vivid examples of when a coding value was applied. The Teamwork discussions were especially good in that they were open to all the team members even if only some of them participated actively and that they could be revisited as and when required, thus allowing to retain consensus on the issues of coding.

3.2.7.3 Collaboration barriers to quantitative data representation

In order to prepare high-quality reports, the project researchers required me as a technical coordinator to produce visualisations and to calculate statistics based on their data annotation. The overall aim of those plots and figures was to provide an effective representation of the key

features of the coded social media posts and the most interesting relationships of those features. While there was a list of key mandatory plot types that each researcher had to include and interpret in their reports, most of the statistics and plots were meant to play an illustrative and supportive function for the researchers' narratives. In most cases, the researchers noticed patterns in the data while coding it and then they would ask me to provide some form of quantitative evidence that could either support or reject their intuition – although, as mentioned in Section 3.2.6.4, occasionally the researchers used the exhaustive exploratory plots to come up with new narrative elements.

To be effective in my role, I had to clearly understand what aspects of the data a researcher's narrative required to represent. Generally, this was not too difficult since the nature of the data and the analysis narrowed down the scope for appropriate representations. In most cases, plots of two types were required: bar graphs that represented a single categorical variable and heat maps that represented two categorical variables[13]. Yet, some confusion still arose from the use of the term *"correlation"*. From the early stages of the project, we all used this term as a convenient jargon word for a relationship between two variables – including categorical ones for which it was not strictly defined. It became apparent that due to differences in the backgrounds of the analysts and myself our understanding of *"to plot a correlation between X and Y"* differed subtly.

For me, due to my background in statistics and data analysis, "correlation" of two variable could be graphically represented *only* by a plot that showed what a change in value of one variable implied for the distribution on another variable. Therefore, the two appropriate plot types were a joint distribution heat map or a normalised heat map that represented conditional distributions of one variable depending on each value of the other variable. However, to my surprise, I repeatedly got requests to plot "correlations" between a *particular value* of one variable versus another variable, such as "Tone: Negative vs. Focus" – or sometimes even between two particular values, such as "Tone: Negative vs. Focus: Britain".

At first, I ignored the specified variable values and produced standard heat maps in response. I mistakenly treated such requests as an attempt at excessive specificity on a researcher's side. However, already during the first round of reporting it became clear that such requests came from most researchers, so I brought the issue up during the subsequent project meeting. As the discussion revealed, the researchers, who all had a background in social sciences or humanities, used the word "correlation" in a looser sense. For them, a correlation of a X with a particular value y of Y simply meant a conditional distribution of X when $Y = y$. As a distribution of only

[13]The main exceptions were the box plots required to support the Instagram analyst who studied the impact of elements of visual content (categorical) on the engagement metrics (continuous).

one variable had to be plotted, bar graphs (with appropriate captions indicating the implied condition) appeared to be a much more effective representation tool. Moreover, a correlation of two particular values thus implied a simple count of the number of instances that satisfied both criteria – it could be represented with a single number and did not require a plot at all!

Interestingly, without being prompted, the researchers never voiced their concerns about the unnecessary heat maps. Most likely, I would have still been producing those unnecessary heat maps until the end of the project had I not raised this issue during a project meeting. This could be partially attributed to the fact that the heat maps still, in principle, contained all the information the researchers wanted to represent. The problem was that they also contained much distracting additional information. Besides, *if* there is a choice between representing the same value with either a height of a bar or an intensity of a colour, the height is easier to for visual inspection and thus should be preferred. Thus, I would speculate that the researchers were silent about this issue for a different reason. For them, I seemed to have authority on the issue of data representation due to my quantitative background. They simply trusted my "choice" of a plot type even if it was not exactly what they had in mind. This exemplifies how important an open discussion is in a team and how valuable it may be to encourage others to voice their concerns even on those matters where they are not the "experts" in the team.

3.2.8 Summary: Key Lessons

The Shakespeare Lives evaluation project was a truly complex social data science endeavour that raised both "hard" methodological issues of making sure that the acquired data and the chosen analysis methods suit the research questions and "soft" organisational issues such as maintaining relationships with the stakeholders and ensuring smooth collaboration among the research team members. The discussion above goes through those issues and elicits insights that should be considered when designing the project management tool for social data science. Even though I will refer back to the detailed accounts of the observations and their analysis given above to inform the tool, it may be worth to quickly reflect on the experience as a whole and elicit some of its key lessons:

1. Stakeholder engagement should be maintained continuously throughout a social data science project. Even if a research team understands the stakeholder interests in principle, there could be differences in details grounded in assumptions and theories rather than in real communication with the stakeholders. Plus, even the correct perceptions of stakeholder demands may diverge from reality over time.

2. When planning data acquisition, equal attention should be given to what the research design

requires, which data actually exist and whether they are available for acquisition. The initial data acquisition plans may change as (a) the bulk of data may come from unexpected sources, (b) different sources may provide varying degrees of access to their data and (c) the planned data acquisition criteria may fail to return sufficient amount of data.

3. Getting to "know" the data through exploratory analysis or in-depth examination of a selected sample is pivotal in social data science. When using novel data sources, a research team may encounter unexpected problems with the data, such as the prevalence of spam tweets, which could otherwise easily be ignored and skew the findings.

4. Researchers should pay specific attention to the derived metrics and their interpretation. Some of the common social data science metrics, such as the sentiment scores, while being intuitive, simply have not been used for long enough and in sufficient variety of contexts. Therefore, the research team may need to adjust their expectations of what a good score level may be.

5. A social data science team may be using third-party tools for acquiring and/or analysing data. Since the terms and conditions of such tools may change over time, researchers may wish to consider how sure they are in their ability to run the same acquisition- and analysis jobs in the future. They should also stay wary of how black-boxed the analyses performed with such tools often are.

6. The choice of specific tools to analyse data with is context-dependent. Even if a particular software solution is normally associated with a particular mode of analysis, project circumstances may override this.

7. If a social data science team predominantly works remotely, the collaboration among its members may be significantly enhanced by using project management platforms where they can share the work artefacts and discuss the project, but also by establishing robust, repeatable collaboration workflows, i.e. the de-facto ways of working.

8. In interdisciplinary teams the members may understand some discipline-specific terms differently. It should be the responsibility of those team members who have a background in that discipline to spot such mismatches and to establish correct shared understanding, since for the rest of the team it may stay an "unknown unknown".

As a final remark, the Shakespeare Lives evaluation project provides an example of doing social data science in a more-or-less traditional stand-alone research setting: while the research was applied and was meant to really shape actions of its key stakeholders (the British Council),

its outputs were presented in batches at three discrete points in time and the produced recommendations had a strategic outlook. This does not affect the value of this case study, but calls for analysing a case where the research process would be more closely integrated into the wider context of a business endeavour and would inform its operationally (cf. Davenport and Patil, 2012; Davenport et al., 2012). The following section does exactly that, as it discusses production of a data-driven radio show.

3.3 Production of Hit List Show for BBC Radio 5 live

This section describes a social data science project in data-driven content production. The project aimed to deliver a weekly radio show for BBC Radio 5 live that presented a the most popular stories of the week, based on an analysis of online data. The show was produced by the Wire Free Production company in collaboration with a team of data scientists from the University of St Andrews and the University of Warwick. This section first introduces and describes the project and then discusses the issues that the team faced while working on the project.

3.3.1 Case Overview: BBC Radio 5 live Hit List Show

The Hit List[14] show was a BBC 5 live radio programme that presented a "rundown of the top 40 news, politics, sport and showbiz stories of the week that are making the biggest impact across social media and online" (British Broadcasting Corporation, nd) among the online audience in the UK. The content of each rundown was based on the outcomes of a weekly data analysis exercise in acquiring and aggregating data from a number of social media and other online data platforms. Aside from occasional breaks, the show aired every Sunday evening from November 2014 to November 2016. Two Christmas specials – 2015 and 2016 – presented the charts of the most popular topics for their respective years.

A normal episode lasted for 2.5 hours, although on several occasions shorter episodes (either a top 30 or a top 20) aired due to the BBC 5 live's schedule. Given the fact that BBC 5 live Radio aired a short hourly news block, the effective length of the show was less and allowed around 3 minutes per a chart entry on average. Such timescales implied that most of the chart entries were treated very briefly with the show's presenter – Emma Barnett – providing a short summary of the story. Other chart entries – internally called *features* – were developed in more detail. In most cases, Emma conducted phone interviews with people connected to the story covered or with relevant expertise. Additionally, Emma discussed the features with guest stars who would be in the studio for the duration of the whole show. Occasionally, Emma played back excerpts of relevant audio material. The feature stories were spread throughout the show, although their frequency did increase towards the end of each programme, i.e. in the higher chart positions.

The process of show production was evolving throughout the course of the show with the most drastic changes happening in the first year. These changes concerned the particular platforms that data were collected from and the particular ways the data from each platform were treated and aggregated. However, the overall workflow of an episode production remained constant.

[14] While the show's title may be spelled as a single word *Hitlist*, I adopt the spelling suggested by the BBC's website (British Broadcasting Corporation, nd).

Each Friday, the team's lead data analyst (hereinafter *the analyst*) collated data that had been already acquired for that week and produced the preliminary chart. The preliminary chart was discussed by the production team in a meeting and topics were assigned to individual journalists to work on. On Sunday, the cycle of data analysis, team meeting and pre-production repeated to create the final version of chart. The following paragraphs outline the contents of each stage.

Data analysis. The analyst had to start very early each Friday (around 6am) to prepare the chart on time. While the specific details of the involved work can be more productively discussed in conjunction with the issues that arose in the process, the key tasks of the analysis were the following:

- determine what the most popular individual entities were on each platform in the UK that week: hashtags for Twitter, posts for Facebook, videos for YouTube or search terms for Google Trends;
- filter out those entities that were topical rather than background noise (e.g. computer gaming content on YouTube, spam on Twitter);
- aggregate the individual entities into clusters each representing a single topics;
- based on the popularity scores on individual entities within the cluster, order the topical clusters into a hit list.

After completing the task, the analyst disseminated the derived chart of topics to all the team members. The chart was implemented as a Google Spreadsheet. Since the analyst worked remotely, the dissemination happened via e-mail.

On Sunday the analyst could afford to start around 10 am. They updated the hit list with the data from Friday and Saturday and disseminated the final chart. The differences between the Friday and the Sunday versions of a chart were normally small enough not to cause major issues.

Team meetings. The production team consisting of an editor and producers gathered at 10 am each Friday for a round-table meeting to work through the topic chart collated by the analyst. The meetings were scheduled to last for 2 hours, but occasionally took just over an hour. The producers split the weekly topics among themselves and each producer would do some quick preliminary research on the stories associated with their share of topics. During the rest of the meeting, the team discussed each topic and made key production decisions:

- which chart entries would be developed into the feature stories;

- the angle under which each topic would be covered and the implications in terms of interviewees, studio guests and further research to be done;
- responsibilities for pre-production of the topics.

These discussions were moderated by the show's editor. There were two interchanging people on the team who worked in this capacity. When neither of them was available, a senior producer took over this role. Throughout the discussion, the editor was standing in front of a flip chart and was documenting the decisions made on each topic. Although the decisions were made collectively, the editor had the final say on all of them, including each producer's pre-production responsibilities. Interestingly, one of the producers on the team was the CEO of WireFree and thus technically the boss of the team, but he happily delegated his power to an editor since those were responsible for finalising the content of the forthcoming show.

On Sunday, the teem meetings took place at the BBC's New Broadcasting House in London. The team started significantly later than on Friday – normally at 1 pm – and usually took no more than half-an-hour for a meeting, as only few changes to a weekly list had to be discussed. In those meetings, the producers would jump straight into discussion of their progress in regard to the tasks allocated on Friday, examine the updated chart produced by the analyst and make final decisions on the content of the hit list and on the featured stories.

Pre-production. The pre-production took the rest of the production team's day. The work on pre-production was carried out individually by each team member with minimal interaction among the producers. The lion's share of the work was in securing guests and interviewees for the feature stories. The producers made use of their extensive contact networks and searched for available contact details to reach out for potential guests and make the necessary arrangements. Other forms of pre-production included preparation of playbacks – e.g. cutting fragments of audio from YouTube videos or pre-recording audio interviews with the guests who would be unavailable to reach during the show – and writing scripts (text to read during the show) for the presenter.

Pre-production work did not differ much on Friday and Sunday, other than that on Sunday it was occasionally a bit more tense. While on Friday each producer could work at their own pace, Sunday confronted them with, as the show went on air at 7.30 pm. The producers also had to make sure to submit their scripts, so that the editor cast an eye on the individual scripts and to put them into the overall script of the upcoming show. Arguably, it was those two factors that made all the team members to work on Sunday together in one production space in the BBC Broadcasting House, whilst on Friday some of the producers consistently chose to work from

home, though office space was available to them in New Broadcasting House.

Overall, the process of working on each Hit List episode was an involved one with the analyst and the producers having to make numerous decisions on how to deal with the weekly acquired data, how to interpret them and how to use those interpretations for the benefit of show production. Those decisions presented an interesting object of study for this thesis and the weekly repetition of the production workflow made studying them convenient. Below, I outline how I approached the fieldwork for this case study.

3.3.2 Fieldwork Methods

The methodology of studying the production of the Hit List radio show is similar to that of the previously discussed Shakespeare Lives case study, as it is rooted in participant observation of the studied project. However, there are some important distinctions caused by the circumstances.

I did not have an opportunity to join the Hit List team until 13 months into the show. By then, a lot of the most influential decisions on how the online data that informed the charts were treated had already been made and the data analysis job had been significantly streamlined. Moreover, I could not join analyst in their weekly work – as mentioned above, the analyst worked remotely and not in business hours. On the other hand, I could join the production team meetings and observe the pre-production work when the producers worked at New Broadcasting House. The fieldwork methods employed were designed to make the best use of the circumstances outlined above. The table below summarises the activities that I undertook while studying the project:

Activity	In co-presence with...	Involved active participation?	Periodicity
Participant observation of production work	Production team	Occasionally	Weekly
Data consultancy	Production team	Yes	Weekly
Post-hoc interviews	Two producers	No	One-off
Review of data analysis methodology	Lead analyst	Yes	Sporadic
Modelling data processing pipelines and code review	Lead analyst	Yes	One-off
Preparing the topic chart for a Christmas special show	Lead analyst	Yes	One-off
Designing improvement for the use of machine learning	Data science team	Yes	One-off

Table 3.2: Fieldwork activities undertaken to study the production of the Hit List show.

From January to November 2016 I was present at most Friday meetings of the production team. Those were the key points in time when I could *observe the producers making sense of the data* presented by the analyst. While most of my observations were passive, sometimes I took on a *more active role* in the production work by taking on a small share of topics to do background research. This helped both with understanding the mindset of a producer and with establishing good relationships with the team, since some of my production suggestions were developed into feature stories on the show. On a couple of occasions the lead data analyst managed to make their way to London so we could observe the production work together and reflect on that cooperatively.

In addition to observing and participating in the production work, I took on a role of *data consultant*. If the team had questions or concerns about particular entries in a weekly chart, they asked me to investigate those either by looking deeper into the data and/or by communicating with the analyst via email or chat software. This role allowed me to make my presence to be continuously helpful for the production team and thus to further strengthen my relationship with them. Additionally, by studying the data and having discussions with the analyst I significantly deepened my insight into the data analysis, its strengths and limitations.

After the end of the Hit List run, I *interviewed two producers* on the team who participated in Hit List production throughout the two years of the show's course. These interviews helped to calibrate some of the fieldwork notes, to elicit the producers' own perceptions of the data and the analysis and also to provide context to the observations – e.g. in how the later months of working on the Hit List differed from the earlier ones and how the experience differed from other projects they had participated in.

The analyst and I also had a number of remote and face-to-face meetings in which we *reviewed the methodology of data acquisition, cleaning, quality assurance and analysis*. We discussed possible improvements to the methodology and *reviewed the analysis code*. For example, once we discovered and fixed a small bug in the acquisition code for Facebook data. Some of those sessions were aligned to coincide with a particular analysis challenge that the analyst faced – for example, *charting the topics in a Christmas special episode*. Furthermore, the analyst provided me with additional resources to study their work even deeper. We *graphically modelled* the data processing pipelines for three out of the four online platforms used in data collection. In addition, the analyst gave me access to all the code they had developed for the Hit List so I could review it at my own pace.

While the wider data science team was involved in the production of the show only sporadically, they held one collective meeting towards the end of the show run. The meeting had a strategic

outlook and mostly generated suggestions for future runs of the Hit List that could have happened if the show had got recommissioned. In that meeting, I *actively participated in brainstorming changes to the design of the new machine learning algorithms for automated topic detection in the data*. Those changes were supposed to improve the algorithm performance and make them more useful for the analyst.

3.3.3 Findings 1. Data Acquisition and Pre-processing: Linking Large Datasets

boyd and Crawford (2012) define "Big Data" through three pillars: technology, analysis and mythology. While in this thesis I stray from using Big Data terminology, preferring social data science instead, it is noteworthy how well the three-pillar framework matches some of the key issues involved in the production of the Hit List. To start with, the technological pillar is defined as "maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets." (boyd and Crawford, 2012, p. 663). Linking and comparing datasets lay at the heart of the process of composing the Hit List chart, as the analyst collated data from different online platforms. This is by no means an exclusive property of the Hit List case compared to the other cases studied in this thesis. In the previously discussed "Shakespeare Lives" project our team also dealt with data from Facebook, Twitter, VK.com and Sina Weibo. However, two particular aspects are specific to the Hit List case:

1. The data from different platforms were "pre-cooked" to different degrees.

2. The analyst had to collate all the data into one and only one hit list.

Both issues are discussed in detail below.

3.3.3.1 How raw the "raw data" were

After the first several months of the show's run, the list of platforms was firmly established as Twitter, Facebook, Google Trends and YouTube. The only mode of accessing data from those platforms that was realistic given the budget constraints and the need to produce the show weekly was to use automated data acquisition jobs that connected to the platforms' APIs or feeds to obtain the platforms' "raw" data. However, as Gitelman 2013, p. 3 notes, there is no such thing as truly raw data, since "data need to be imagined as data to exist and function as such, and the imagination of data entails an interpretive base". While all the Hit List data captured one underlying phenomena – a user's interaction with an online platform – the way this essence was represented in the acquired data varied both due to the specifies of the data sources and the

decisions of the analyst. As a result, at the very moment of acquisition the data from different platforms were *"pre-cooked"* to a varying degree.

Twitter data were arguably the rawest. The analyst employed the Twitter Streaming API's Sample endpoint to continuously acquire a 1% random sample of all the published tweets in the world in real time. Such data were potentially the richest, however they also required the most initial treatment from the analyst before the core analysis tasks (identification of which narrow themes in the data were topical/news-worthy and how wide the boundaries of an individual theme were) became at all possible. First of all, the data coming from parts of the world other than the UK had to be filtered out, which is discussed in detail in Section 3.3.3.3. Second, the individual tweets had to be somehow thematically aggregated. The most straightforward way to do that was through identifying the most commonly used hashtags and collating the tweet counts for those that belonged to one topic. Since multiple hashtags were often used in one tweet, the analyst first derived the list of the most common hashtags and then, for each of those common hashtags, a list of their co-occurring hashtags. The analyst used their judgement to pick-and-choose from these lists, thus forming topical hashtag clusters.

In the case of YouTube, the analyst could make use of quite significant data "pre-cooking" on the platform's side thanks to YouTube's automatically generated and continuously updated playlists of most popular videos in each country, including the UK[15]. While YouTube did not openly state how often the UK playlist was updated and how long it was, through trial-and-error the analyst determined that the updates happened approximately every 20 minutes and that the playlist always consisted of 200 videos. Thus, the YouTube data collection job made calls to the YouTube API 3 times an hour throughout the week to access the current contents and order of this playlist. Such use of the playlist made the acquired data pre-cooked on several levels compared to Twitter:

- Out-of-the-box, the data reflected the interests specifically of the UK public.
- Instead of coming at the level of individual user interactions (posting a tweet), YouTube data came at the level of individual objects (videos) with which the interactions (watching, liking, commenting, etc.) could happen. Therefore, no additional analysis stage of reducing individual interactions to thematic objects (analogous to reducing tweets to their hashtags) was required.
- The order of videos in each individual playlist gave the analyst a pre-cooked popularity score for each video. Instead of deriving their own popularity measure (e.g. some integrated

[15] At the moment of this writing, UK's playlist can be found at `https://www.youtube.com/playlist?list=PL-DfNcB3lim9IZmUXEjE1Ov0Ir1NDa3Yr`

rating that would take into account views, likes, comments and other metrics), the analyst could simply assign a high score to the videos at the top of the playlist and a low score to the videos at the bottom of the playlist. The procedure that YouTube used to order its auto-generated playlists was completely black-boxed, however it was not that much of a concern in the case of the Hit List. Moreover, YouTube's playlist ordering was most certainly better informed than any measure the analyst could design, since the publicly available data of video's popularity were presumably not as rich as YouTube's internal data. The use of this pre-cooked popularity score avoided repeated additional API calls to acquire details on each individual video, which might have led to rate-limiting. The considerations of what data to use therefore were influenced by other factors, to do with the proprietary nature of the data.

The data from Google Trends were acquired in a similar manner – i.e. repeatedly with a 20 minute interval – from the Google Trends Atom feed. In many ways these data were similar to YouTube data (presumably since both platforms belonged to the Google product ecosystem). The data represented a list of the most popular Google search terms in the UK at the moment that could be treated in a similar manner to the YouTube playlist. The only difference was that the Google Trends lists of search terms were sorted by only one parameter – the total number of searches for a term within the reported time frame – and that the data did include an approximate number of the search requests, for example "100,000+" or "5,000+" with larger numbers having lower granularity. Having to aggregate such estimates rather than exact values was a typical example of a limitation associated with dealing with proprietary data. However, since such aggregates provided insight into at least the order of magnitude to the search term popularity, using those still gave higher precision than only using their chart positions in the Atom feed. The subsequent aggregation of weekly data for both YouTube and Google Trends boiled down to summation of the chosen popularity scores for each video / search term across all the considered scrapes.

The Facebook data were almost "well-done" immediately at the point of acquisition. In principle, the Facebook Graph API allowed to acquire data on an arbitrary level of granularity – e.g., if there had been such a desire, the analyst could acquire detailed data on each individual "like". However, as the experiments revealed to the analyst, Facebook put severe restrictions on the number of requests that could be made to the API, as the speed with which Facebook returned responses to the API requests fell dramatically over time. Because of that and a lack of convenient methods to collect a random sample of the relevant Facebook data, the analyst had to come up with an acquisition approach that would be very selective both in terms of the particular Facebook pages – i.e. either individual user pages or organisational public pages – and in terms of the acquired data types.

The chosen approach was to limit the acquisition to weekly posts from the public pages of the UK news media organisations with substantial following on Facebook and to use post-level engagement metrics – comments, likes and shares. This approach involved a noteworthy trade-off: on the one hand, it went a bit against the motto of the show as it did not expand the show's agenda beyond the items already reported in mainstream news media. On the other hand, the derived Facebook data were very convenient to work with as by their very nature they represented almost exclusively news-worthy topics judged to be interesting to the UK public. Minimal moderation from the analysts side was required. When comparing this back to the complex process of making the Twitter data workable, it becomes self-evident that calling data coming from these two platforms equally "raw" would be a drastic oversimplification.

3.3.3.2 Aggregating data across platforms

While the discussion above shows the complexity of bringing the data to the state when the popularity of the represented topics could be assessed within each platform, turning those platform-wise popularity estimations into a coherent integrated rating was a significant issue on its own. The data from different platforms represented qualitatively different interactions between the users and the topics and were also affected by the relative popularity of the platforms and the difference in their demographic profiles (Duggan and Brenner, 2013). It is non-trivial to judge how much a single Facebook comment in search queries or in tweets. Moreover, the YouTube score, as discussed above, was based on relative positions of a video in an auto-generated playlist rather than on raw counts of underlying user interactions.

The approach that the data science team agreed upon was making separate charts for individual platforms and then aggregating those via assigning each platform a weight. Thus, the 40 most popular topics on each platform received a platform-specific score distributed from 40 to 1 inversely to the topic's position in the platform chart. A weighted sum of those scores across all platforms then gave a topic's overall score.

This approach had profound consequences for the contents of the show. First, it allowed a good degree of variety in the covered topics. Each platform had its own characteristic trending content – partially due to the platform mechanics and the demographics of their user-base, partially due to the way we collected the data from each platform (cf. Ruths and Pfeffer, 2014). For example, YouTube appeared to be prone to carrying "viral", entertaining content, while Facebook, by the construction of the data acquisition process, was "newsy". With the approach taken, the most popular content from each platform was presented somewhere in the chart even in those weeks when that platform had not gained high absolute levels of interaction. On the other hand, a platform-specific topic, no matter how overwhelmingly popular it was within one platform,

could never take a top spot from a theme that was trending across multiple data sources. This sometimes led to a more predictable top-5/top-10 (with Brexit and the US Presidential Elections being at the top of the list for around half of 2016), but gave us additional confidence in the top slots of the chart.

However, arguably even more important consequence of the approach taken to link the platform-specific datasets was the ability to weight the platforms differently. The weights fluctuated during the first year of the programme, but from August 2015 to the end of the show they have remained constant. The key factors behind the chosen weights were:

- *A strive for a diverse and balanced list of topics.* Some of the key dimensions to balance were (a) uniqueness of the topics in a weekly chart vs. coverage of the topics trending across the mainstream media, (b) UK vs. international focus, (c) hard news vs. entertainment. The degree to which the balance was achieved was assessed mainly by the show producers, for whom this balance was a major selling point of the show. As one of the producers said in an interview, "the breadth of the stories [...] from either side of the spectrum and everything in between was vital to what made [the Hitlist] so good". Moreover, as her colleague observed:

  > "I think [the Hit List] is really reflective of how people digest news. So many people actually just digest news through social media, so they are looking at one second a cute panda and the next second Donald Trump. [... Traditional] news, they haven't caught up with this; that's why this is so important."

- *Striving for a workable list.* One aim in compiling the list was to allow the show producers to give the topics a proper journalistic treatment. While the producers were up for the challenge that the data-driven reporting brought and were happy to study topics they would not normally come across in their work, some topics lacked in substance beyond repair by journalistic work. YouTube, as the least event-driven and the most content-driven platform, contributed a lion share to this problem, which naturally led to it being the lowest rated platform. In addition to using the weightings, informal rules emerged that aimed to exclude content that did not fit the definition of "news" for a general audience. Examples were music videos or gaming videos (see Section 3.3.4 for a more detailed discussion of content exclusion).

- *The level of confidence in and experience with processing data from various platform.* At the very start of the show Twitter was the only data source. While its data were in a sense the most problematic to deal with, the data science team members (including the lead analyst) had had the most previous experience in dealing with Twitter data out of all online platforms. For these reasons, Twitter stayed the highest weighted platform for the course of the show.

By contrast, Facebook was initially added with a low weight, but after several months, when its high news-value and important role in counter-balancing data from other platforms had been robustly established, it was advanced to the same highest weight as Twitter.

It is worth noting that the motivation behind the selected weights was very practical and pragmatic. The weights did not necessarily reflect our perception of the relative importance of the selected platforms, or our understanding of the overall volumes of interaction on them. Rather than that, they were used to support the end-goal of the project: production of an interesting and varied weekly news chart for live radio that would resonate with the audience. This may seem a bit counter-intuitive given the show's data-driven, evidence-based nature, but arguably is a common goal for all such attempts at creating "trending" features – for example, in the case of the aforementioned YouTube playlist of the most popular UK videos, it is hard to imagine that Google compiles it for any other reason than engaging audiences.

3.3.3.3 Filtering UK data

The discussion above suggests that most of the data processing that had to be performed beyond the straight-forward automated data aggregation was done manually by the analyst who exercised their judgement on the topic boundaries and news-worthiness. However, it has also been mentioned that the Twitter data were acquired in the form of a real-time 1% random sample. This motivated the use of machine learning to filter out the tweets that did not come from the UK. Indeed, when charting the topics discussed on Twitter, the analyst could not afford to only rely on the tweets that contained geolocation data, as such tweets were extremely sparse. The geolocation codes are contained in only about 5% of tweets (Graham et al., 2014) – and that would have been 5% of a 1%. Hence, for the vast majority of data we had to infer their country of origin using a classification algorithm. The analyst employed a Bayesian classifier that had been implemented by a different data scientist on the team and had been trained on the tweets that did have geocodes. A revised version of the classifier is discussed by Zubiaga et al. (2017). The version of the classifier employed in the production of the Hit List reported on precision and recall at 85% and 68% respectively, which implied an expected level of 15% false positives in the filtered data. It is worth examining how the application of an algorithm with known limitations affected the chart itself, the team's perception of that chart and how it was dealt with.

While the filtering algorithm definitely allowed for a collation of a much more UK-centric list of topics, it systematically left in a loosely defined set of topics (expressed as hashtags) that the Hit List team strongly suspected to be false positives. The suspicion was rooted in (a) the absence of these topics in the charts of the other individual platforms (b) the Hit List team's expectations and perceptions of what might be of interest for the UK public on Twitter and beyond and (c)

common features shared by many of those topics. Those topics tended to be US-centric, which could be explained by the fact the the US sector of Twitter was by far the largest and that the majority of British and American tweets shared a common language, which could lead to further confusion for the classifier. Some of those topics were left in the chart (e.g. the highly topical political ones around the US presidential election and the Black Lives Matter movement), while others (e.g. related to some of the US-specific TV talk shows) were eliminated manually by the analyst. Interestingly, if the analyst had not manually discarded those topics, some of them would have consistently charted for consecutive weeks. Such repeating irrelevant chart entries could have been specifically annoying to the show's audience.

A closely related concern was in regard to the low recall rate of the algorithm. The time- and resource constraints of the Hit List production did not allow for studying the unfiltered version of a chart, so while the analyst did their best to deal with false positives, they were ill-equipped against false negatives. In principle, since the hashtags/topics were not used as features in the country classifier, it was not completely unreasonable to assume that a probability of a UK-originating tweet to be discarded was independent from the tweet's topic – however, this assumption never was actually tested. Potentially, there could have been topics popular in the UK that were pushed down the list significantly more than others. Since the analyst practically never looked further down the list of hashtags than the top 100, some of those topics could have never even got a chance to be noticed. Yet, the Twitter charts were sufficiently varied and by far and large did not contradict the expectations of the analyst and the production team, so the pragmatic reasons to investigate this issue were not worth the required resource investment.

The discussions with the analyst also indicate that over time they got accustomed to the systematic biases of the country classifier (or, more precisely, what the Hit List team perceived as those systematic biases). In addition to the machine-learned classifier, the analyst trained themselves to be a better "human classifier". Some seemingly odd topics – or rather types of topics – appeared in the data so *consistently*, that over time treating them became routine. The analyst and the production team would agree to either systematically leave in or discard those topics from week to week. By contrast, the questionable topics that were presented in the data as a one-off on a particular week could cause hesitation. Most often, the analyst would leave them in for the production team to decide. The production team would normally come to a definitive conclusion on their own or after further consulting with me and, if required, the analyst. On several occasions, making a judgement call was extremely tough. For example, one week an unexpected hashtag *#purple* trended in the data on Friday. The team went through some struggles before identifying that the hashtag corresponded to then-premiering movie "Purple" about the Ghanaian music scene. Lack of coverage of the movie in the mainstream media and

the perceived unexpectedness of the topic raised eyebrows during the team meeting. However, as Ghana-related topics were not the "usual suspects" for geographical misclassification, we did not feel we had enough grounds to exclude the topic from the chart. Ultimately, the topic appeared to be below the top-40 on Sunday, thus freeing us from the need to decide its fate.

Interestingly, the use of a machine learning algorithm to filter data potentially affected the "human algorithmic" treatment of some topics. Indeed, as the analyst worked with the data that had been pre-filtered by a classifier, they were arguably a bit more reserved in their own data manipulations. The aforementioned topic of the US elections provides an example. It would not be unreasonable to assume that its consistent overwhelming popularity according to our data was at least *partially* due to misclassification. In fact, this suspicion was shared by the analyst, the production team and myself and was amplified by the fact that the US elections were consistently charting as the top-1 topic on Twitter for about half of 2016. To remedy this, we could have decided that the analyst should penalise the US-related topics when constructing the chart. However, as we knew that (a) the UK population talked about the major US-related topics anyway and (b) crucially, the data had already been pre-filtered by a trained classifier, the decision to penalise these stories was never made and the analyst did not adjust the scores of the US-related hashtags.

3.3.4 Findings 2. Data Analysis: Identifying Hit List Chart Entries

Determining which data represented topical stories and which boundaries the respective topics had constituted the bulk of the analyst's work. This task also involved the greatest amount of moderation by the production team. In this sense, it was arguably the most cooperative part of Hit List production. Each week, when distributing a new chart to the team via email, the analyst reported on all the core decisions that had went into the chart collation, so that the team could question those and potentially reverse them. Some decisions, as already indicated in Filtering UK data, could be applied to data consistently from week to week and thus did not require reporting every time – however they did require initial coordination and the producers' approval. Finally, there were cases when the analyst did not notice a reason to discarding a topic or for aggregating / splitting topics, in which case it was the production team's judgement that would drive the decision.

Cleaning signal from noise. While most of the weekly data acquired for the Hit List production indeed corresponded to a particular story happening that week, some of them were what the production team called "background noise". The noise was often caused either by (a) natural routine activities of platform users and (b) artificial boosting of particular types

of user-interactions. Gaming-related YouTube videos were a good example of the (a)-type noise. Those videos are genuinely popular and constitute a large portion of YouTube's content consumption. In December 2014, it was reported that 15% of all YouTube videos showed the process of playing video games (Marshall, 2014). At least 6 out of 10 YouTube channels that have been recently reported to be the most subscribed-to in the world are dedicated to gaming (Fitz-Gerald and Butkovic, 2018). It is not surprising that the game-play videos frequented the unfiltered YouTube charts. As these videos usually did not provide anything for the production team to report on, they were routinely discarded from the very start of using YouTube data.

As an example of the (b)-type noise, Google Doodles skewed Google Trends data. A Google Doodles typically is an interactive short animation that Google puts on its search engine's home web-page. These animations often correspond to an anniversary of an event that Google's staff consider significant, but that would often not be of major interest to the general public otherwise. A recent example is a Doodle that celebrates the 62nd birthday of an Ebola-fighting physician Dr. Stella Adadevoh[16]. A day's Google Doodle is displayed to everyone who visits Google's front page. Normally, if a user clicks on a day's Doodle they are redirected to a page displaying Google search results for the person or the event portrayed in the Doodle. These searches, while arguably being often spontaneous or even accidental, dramatically boost the respective search term's positions in Google Trends data.

The decision that the analyst should exclude the data on the Google Doodles searches from consideration was agreed at the early stages of the project. Interestingly, in one of the last weeks of the Hit List airtime, the analyst accidentally left a Doodle-related data entry in when compiling a Friday version of a chart. That Doodle was dedicated to an anniversary of the first edition of Michael Ende's best-selling fantasy novel "The Neverending Story". One of the team's producers was so touched by the apparent public response to the anniversary that they suggested to develop it into one of the week's feature stories. In turn, I did happen to notice the source of the anniversary's "popularity" and, as a data consultant, felt responsible to point it out to the team. To my great surprise, it appeared that no one on the production team even remembered the decision to discard Doodle-related data in the first place. And, since the story was very appealing to the team as a whole anyway, it was left on the chart regardless.

The above-discussed episode highlights two important points. First, the long-term decisions on how the analyst should treat the data were essentially heuristics. They did not warrant to always provide the best result – rather than that, they aimed to secure *sufficiently good results on a continuous basis* and to also be workable. Second, while implementation (and control of implementation) of those decisions was routine for the analyst and for me and therefore we were

[16] https://www.google.com/doodles/stella-adadevohs-62nd-birthday

fully aware of those decisions, over time such decisions became effectively black-boxed for the production team, even though they had participated in making them.

To finish the discussion of the types of background noise in the data, it is worth noting that some of the discarded themes in the data were background noise of types (a) and (b) simultaneously. The prominence of hashtags related to popular boy-bands, especially "One Direction", in Twitter data was an example of that. On the one hand, the numerous fans genuinely discussed "One Direction". As such, the hashtags in the data were a by-product of normal routine of UK Twitter users. However, one aspect of the Twitter interactions around "One-Direction" and other acts popular among similar demographic groups provided an additional boost specifically to the use of hashtags: *voting*. One of the producers of Hit List recalls this issue in an interview:

> "Some voting things, MTV music awards and things like that, like American music awards [were] some things I clearly felt were there because people were voting for them, and that was like teenagers going mad and clicking things a million times."

Indeed, music award pages, band fan clubs and other public pages on Twitter dedicated to entertainment for teenagers frequently encouraged their followers to vote for one act or another. An eligible vote usually had to come in the form of a tweet with a suggested hashtag. Thus, even those fans who would otherwise prefer not to use hashtags (or would forget to use them), were helped to reconsider. While over time the analyst did improve in identifying and discarding poll-provoked data consistently, in the early weeks of Hit List those data did pose a challenge.

3.3.4.1 Questioning essence of news

As the reader can conclude from the discussion above, at least a conceptual understanding of what constituted "background noise" in the data was established quite early in the cycle of Hit List production. However, this did not mean that anything that was *not* background noise actually had any reporting value and was genuinely news-worthy. And even if there was some "story" attached to a chart entry, it was not necessarily appropriate for nationwide broadcasting. This all posed the question of what 'news' was and whether something had to be news in the first place to appear on Hit List. In practice, this question could not be answered definitely and had to be dealt with on a case-by-case basis with the answer depending on the particulars of the story, on its wider context and partially even on producers' own research and journalistic treatment.

For example, once a viral video that displayed a man sexually pleasing himself with a McDonald's McChicken burger trended in the data. The video's gross content and apparent lack of substantial story made it an almost sure candidate for exclusion from the chart. However, it was decided

during the production team meeting to do additional research into the story to make sure that pulling the plug was the right thing to do. This research revealed a very interesting background story to the video's popularity (Thielman, 2016). Not long before the release of the video, Facebook changed the way the "Facebook Trending" feature was compiled. The company switched from manual curation of data (not unlike the one employed for the Hit List production) to an entirely automated machine-learning solution. The company thus responded to a public backlash, as its editorial had been accused of left-wing bias. The change to automatic generation of "Facebook Trending" subsequently brought a lot of problems, among which was the failure to filter out the McChicken story. This exposed the video to a large segment of Facebook audience, providing a positive feedback loop for its popularity. The topic of algorithmic failure in news curation was so exciting for the Hit List production team that the McChicken story was not only left in the chart, but also developed into a feature.

Sometimes online content that attracted user interactions was not new by itself, thus begging the question of whether there was a fresh story attached to those interactions. For example, a YouTube trailer of a movie "Straight Outta Compton" trended in September 2016. Trailers frequently charted high in YouTube data, however this particular trailer was for a movie that had been more than a year old. As a data consultant, I was asked to investigate the case. The real story that caused the surge in the trailer views appeared to be the death of Jerry Heller – an American music manager who had worked with many rap artists and producers whose life stories were captured in that movie. By contrast, in other cases, the online content that was seemingly new after further investigation appeared to be merely a new upload of old materials. Such content often got popular not because of new story developments but because some of the audience had missed it the first time. Yet again, this was a common case with YouTube videos – for example, a video of a woman having a mental breakdown in public transport that charted once in 2016 was a new upload of a video from 2011. Facebook pages of news media, especially tabloids, also were noticed to reprise old content, presumably for click-bait. For example, the Daily Mail once "reported" on an old video of a school boy winning a fight with a bully. In most cases these uploads of old content were discarded by the production team, however it was always a decision made with discretion.

3.3.4.2 Defining topic boundaries

Similarly to deciding whether to discard or to consider a particular story, the decisions on the topic boundaries could be either strategic or ad-hoc. As an example of a case that required to be handled strategically, during the English Premiere League football season, discussions about the games and about the football clubs involved were consistently trending in the data. That posed a challenge. While the ongoing developments within the Premier League were not

seen as non-topical and thus completely ignored, most of the time they did not involve any specific "stories" to report that would go beyond the details of particular games. Furthermore, BBC 5live aired a lot of sport-related programmes anyway; some of them were right next to Hit List in the programme schedule. Therefore, featuring sport games on Hit List would have sounded redundant. From this perspective, it might have seemed reasonable to collate all the football-related data into one large topic and to talk about it only once during the show. However, such a topic would have consistently been at the top of the chart and thus it would have suggested a full feature coverage. By contrast, treating each football team as a topic of its own occupied more positions in the chart – but those were normally lower than top-10, thus allowing to give them only a brief mention and to focus the discussion on other stories.

While each football team was treated as a topic in its own, the online discussions often focused on particular *games*, thus involving two teams at once. Some of the most highly trending hashtags in Twitter data had the form of *#[team]vs[team]*, e.g. *#MUNvsMCI* for a Manchester derby. The analyst had to decide how to deal with such cases. It was decided early in the programme that the analyst would split the points for "collective" hashtags between the mentioned teams. The splitting could have been either weighted (for example, proportionate to the number of isolated mentions of each team) or unweighted, i.e. half-and-half. Although both versions were not difficult to implement, the analyst went with the unweighted split. This helped to further keep the football lower on the chart as the more popular team could not get a boost from more split points in its favour. Over time, the practice of half-and-half split became an analysis routine and started to be applied to other cases when splitting was required. An example of such event was the Iowa caucuses – a political event in the state of Iowa that essentially represented the first act of primary (inter-party) elections for both the Democratic and the Republican parties. The caucuses provoked the use of generic hashtags for two distinct stories, one about each of the two parties.

On several occasions, a connection between two topics became evident only as a result of in-depth examination of the related stories. This was especially true for one-off topics that required ad-hoc treatment. For example, one week Facebook data showed a high level engagement with the BBC's article profiling Tammy Saunders who had suffered a partial loss of her face. The same week, a then-recent episode of "The Undatables" – "a dating programme for people with challenging conditions" (Hawkins, 2016) – charted on several platforms due to the appearance of a popular rugby player. At first glance, the stories seemed to bear no relationship. Therefore, the analyst, who had limited time to work through the story details, placed them as two different entries on the chart. Luckily, the production team became really interested in Tammy's story. By digging deeper, they figured out that the reason her story had got spotlight in the first place

was her participation in one of the previous episodes of "The Undatables". This discovered link allowed to strengthen the coverage of Tammy's case by contextualising it in the broader discussion around the TV programme.

3.3.4.3 Machine learning for topic classification

Since defining topic boundaries was such a big task, an attempt to assist the analyst by the means of employing a machine learning algorithm was taken. The same data scientist who had implemented the Twitter country classifier also implemented an algorithm for clustering data basing on the vocabulary employed in their textual fields. The underlying assumption was that different topics required different words to be discussed, and therefore the resulting data clusters would correspond to individual topics. The clustering outcomes were thought to be used as a starting point for the analyst's work, who would then tweak the derived clusters as required.

The clustering algorithm was tested by the analyst in his work on data from Google Trends. This platform was chosen because of the richness of the textual descriptive metadata that Google returned for each search term. The Google Trends data included a short textual description for each search terms plus a list of related news article headlines. Unfortunately, the algorithm performed significantly below the expectations. There was a tendency for a formation of a giant cluster whose member terms were often unrelated, while the rest of the search terms tended to not pair into clusters at all. As a result, while the clustering algorithm continued to be a part of the Google Trends data processing just in case it would flag up a connection between topics that the analyst could miss, the analyst effectively had to aggregate search terms into topic from scratch.

Several factors could have contributed to the algorithm's week performance. Some of those had to do with the specific design decisions that were taken in the implementation of the clustering procedure. For example, it employed a bag-of-words approach to feature extraction (Zhang et al., 2010) – essentially, each of the clustered pieces of text was presented as an unordered set of its individual words weighted by word occurrence in the text. Such a representation significantly reduced the richness of data and could have been too simplistic to capture the similarities and differences between textual descriptions of the compared search terms. The clustering algorithm employed was also designed to make it more likely for bigger clusters to attract new members than for the small ones, which could partially explain the giant cluster anomaly.

However, there was an arguably even larger *conceptual* issue involved. In machine learning terms, the task of identifying topical clusters could only be posed as an unsupervised one. In unsupervised learning classification tasks, the clusters emerge from the data "by themselves" through sheer differences and similarities of the statistical properties of the individual data points.

This is to contrast with supervised learning where the list of the expected data classes is given and the expected statistical properties of those classes can be inferred from the gold standard data. Supervised classifiers are expected to outperform unsupervised ones since the gold standard data can guide them; yet they, by definition, cannot deal with data where the list of classes is not known in advance. Since in the task-at-hand a class represented a topic, the algorithm could only have been supervised if all the same topics had trended each week, which obviously was not the case.

What the unsupervised learning approach missed out – and what the analyst, by contrast, successfully used to their benefit – was that many topics *did* trend each week. The analyst accumulated the knowledge on the reoccurring topics over time to the point of often effortless spotting those topics in a weekly portion of data. Moreover, even dealing with genuinely new topics and search terms became easier for the analyst over time since they got deeper knowledge of the overall news landscape and better understanding of how the data were structured, what to expect of them and how to make sense of them. In a way, while the task of identifying the particular topics continued to be mostly unsupervised for the analyst (bar the "usual suspect" topics), the task of identifying those features of the data that were most helpful to determine new topics was very well supervised by their past experience. Designing a machine learning algorithm that would be able, like the analyst, to use the historical classification examples to guide a future unsupervised clustering, was the only real way to match the analyst's performance to any significant degree. Yet, the complexity of the involved work was way above the time that the machine learning expert on the data science team could afford to invest.

A possible compromise that could have provided at least some real help to the analyst (albeit limited) while not requiring that much effort to implement was to train a traditional supervised classifier to identify the usual suspect topics. In principle, it could even have been used in conjunction with an unsupervised classifier with a condition that the latter would have been applied only to the search terms that did not fall into any of the common categories with a sufficiently high probability. I proposed this solution during a data science team meeting. Unfortunately, that meeting was held closer to the end of the run of the Hit List and was aimed to suggest changes to the production of Hit List that would be implemented if the show was re-commissioned. Since the show was taken off the BBC's schedule, due to external factors, the suggestion did not get a chance to be implemented and tested.

3.3.5 Findings 3. Producers and Data: Establishing Trust

Given how many decisions had to go into collating a weekly Hit List chart and how pragmatic and driven by journalistic interest some of those decisions were, a reader might think that the

production team did not consider communicating data honestly as important as producing an entertaining show. However, in practice, most team members had a very strong ethos of *staying true to the data*.

An example of the team's commitment to the charts was once observed on a Sunday, shortly before an episode of Hit List was scheduled to go on air. That week's data happened to be less interesting than average. The show's presenter noticed that and complained about it to the head of the production team leader, possibly in a hope that some more exciting stories could be injected into the chart. In response, the team leader advised her to reconsider her stance on the chart and to *"treat it like the weather"*, i.e. like something given rather that as something to tweak and adjust.

The strict commitment to the data demonstrates the team's *trust in the data*. The source of this trust varied across different team members. The following discussion shows how the varying attitudes of the team members contributed to trusting the data and what role the form in which the data was presented to the producers played.

3.3.5.1 Attitudes to data

The interviews with the two producers conducted as part of this case study demonstrated very different attitudes to the data that, nevertheless, both resulted in supporting the "staying true to the data" ethos. One of the interviewed producers built her relationship with data based on her understanding of the ethos of working in a production environment. According to her stance, trusting the data was almost a prerequisite for the production work to be carried out.

> "If we are given a chart with the methodology that you [i.e. the analyst and myself as the data consultant] use, we have to trust that your data collection is correct. Because [it is a bad idea to] relate this information to the audience [and] bother them with the science behind. [...] And I am not gonna question how you got there. I'm trusting that you have used your tools and [data science methods to] put it all together correctly for me to make a value of it."

It is worth noting that she also found weekly charts to usually match her own social media experience. The combination of this match and the ethos expressed above made her opinion on sticking to the data very strong. She admitted to "hate" those rare instances when for practical reasons the chart had to be bent and strongly disapproved of "shortcuts" and "quick-fixes" that could ease the producers' lives but would violate the chart.

The other interviewed producer had a more critical outlook on the data. She actively reflected on the analysis involved and had spotted its limitations. For example, she did notice that the charts were often skewed towards the events that had happened early in the week because they had had more time to attract user interactions. She also questioned some aspects of methodology employed by the analyst. For example, she did not particularly agree with the analyst's way of dealing with Facebook data. When ranking the popularity of Facebook posts, the analyst assigned the highest weight to the number of comments as the most complex interaction and the lowest – to the number of likes as the least complex interaction. According to the producer, this was not always the best idea:

> "There's one thing I wasn't that happy with. [...] The weighting [the analyst] gave to things on Facebook that had comments, I thought it was too much. [...] If something was commented on Facebook, it got more points, [...] while I think this whole chart should be about quantity not quality of interaction. [...] I thought that it was a bit disproportionate. Like, sometimes a Facebook video that had been viewed, I don't know, a thousand times was really high in the chart, but actually, it wasn't... You know, sometimes you can get a bit of a gut instinct of this as well? I just didn't feel like it was."

Overall, she confessed to questioning "one or two" stories on the chart each week, although "in massive quantities – not that often". Interestingly, it does not mean that her trust in data was much lower than for the first producer. As the show progressed, it changed from "a bit skeptical" to "critical" to "quite trusting" to "very trusting" – yet, the latter was achieved not though accepting the trust in the data as a necessary prerequisite, but through getting to "know the data and [...] the outliers", so that the data could be efficiently understood and quality assured and the chart could better represent the reality of online discussions around news and topical stories:

> "I think we should be allowed to use our brain to discard those [erroneous chart entries]. [...] You need a brain looking at this, you couldn't just go with this. [...] When you're making something journalistically and for a programme, you should be able to speak to [a] person. [...] It wasn't necessary that [the analyst] would change things [when the production team questioned them], but [the analyst] would explain why the things were the way they were."

This is a very different line of thinking compared to that of the first producer, but it can be seen how it is equally intolerant to making fixes to the chart simply to satisfy the agenda of the production team and to allow for easier production work.

3.3.5.2 Using chart spreadsheets

Another important vehicle of trust into data was the way the weekly charts were distributed by the analyst. As mentioned before, the analyst used to collate a weekly chart in a Google spreadsheet. They would subsequently distribute the chart to the production team by sharing a link to this spreadsheet with the team members. Each spreadsheet had four tabs – one for the overall chart and one for each of the 4 contributing platforms minus Twitter. The Twitter tab was missing since collating a Twitter chat required more work and was done in a separate workbench specifically developed for the job.

The overall chart tab contained a list of approximately the top-100 topics from the data. This was somewhat generous considering that only the top-40 topics were presented in the programme, but sometimes this longer list allowed to produce an interesting "bubbling under" feature – a short segment that occasionally aired at the beginning of the Hit List and talked about one story that the production team particularly liked but that did not make it into the top-40. The longer chart could also sometimes help the producers to predict which of the stories would make it into the top-40 next time. This was especially relevant during the Friday team meetings since it allowed to better prepare for Sunday. Finally, the extended list of stories could also help with trusting the data – if a producer expected some story to be in the top-40 but did not see it there, they could look down the list and make sure that the story was not simply forgotten or discarded by mistake, but rather simply did not make the main chart.

If a chart entry trended on Facebook and/or YouTube, it had a link to the either a relevant popular Facebook post or a YouTube video. The chart entries coming from Twitter or Google Trends, unfortunately, did not have a particular piece of content attached. The more detailed platform-wise tabs allowed to see all the search terms, videos and Facebook posts that went into each charting topics. Both of the interviewed analysts said they had periodically used those to familiarise themselves with the topic and to find inspiration for potential angles of covering each story. Moreover, in the latter months of producing the Hit List, even if the producers did not have time to study those additional tabs themselves, they would periodically ask me to investigate those they had doubts about. As such, the spreadsheet was a very useful tool for data mediation, which of course also helped with establishing trust in the data.

While the spreadsheet interface was quite simplistic, it proved to be sufficient for the vast majority of cases where further examination of data was required. The most clear exception was

examination of Twitter data. The detailed breakdown for those data was available for everyone in the team through a web application that was hosted on the data science team's server. However, since looking at the Twitter breakdown was only required infrequently and since the interface of the web-application was not the most user-friendly, in practice the production team always found it easier to ask either the analyst or me to have a look at it. Other than providing an easier interface to examine Twitter data, the producers only ever mentioned one feature that could have been a useful addition for them to further study the data: timeline graphs of user interactions with various topics. Those could be helpful to search for what was happening *outside* social media just before the peaks of user interactions and potentially to combat the aforementioned beginning-of-the-week bias.

3.3.6 Findings 4. Production Treatment: Turning Data into Radio

The Hit List chart was an unusual format to work with from a production standpoint. First of all, it presented the challenge of turning the stories that the producers would not normally pick up themselves into engaging pieces for radio. Second, the sheer format of a weekly news chart was challenging in its own right. It implied that in 3 days (from Friday to Sunday) the team had to deal with more than 40 stories, some of which were bound to fall off of the final chart, while others were known only on the day the programme went on air. One of the interviewed producers explains that she welcomed this challenge because and that she believed the format of the Hit List was "genius" and unlike anything she had heard of prior to the show. According to her, "it's an easy way to deliver social media area of interest to somebody who doesn't necessarily know enough about it".

The below discussion outlines what the production- and the wider Hit List team do to realise the potential of the show's format and to extract the most journalistic value out of the data.

3.3.6.1 Developing feature stories

The format and the content of the Hit List posed some challenges for developing feature stories. Occasionally, it was difficult to present a story without giving a glimpse of the original piece of online content behind it. As the online content is consumed on screens of computers and mobile devices, it is predominantly visual and not easily transmittable via radio. This was especially problematic for YouTube videos, since their dynamic was not always easy to capture through paraphrasing the content. Quite often, the producer would include an audio snippet from a video into the script. The snippets had to be selected strategically so that they did not require the accompanying picture to be appreciated.

However, according to both interviewed producers, in most cases the process of developing a feature story was not significantly different for Hit List compared to any other radio show they had worked on, as the best way to communicate a story was still to have an excellent guest (either as an actual studio guest or as a telephone interviewee) who would be able to deliver value to the audience in a short time frame allowed for one feature. One of the interviewed producers believed that an initial list of weekly stories and the guests were of "fifty-fifty" importance to how good an episode of the Hit List was deemed to be. The other producer elaborates:

> "[The feature stories] are just stories, this is no[t] difficult. The transference of information to radio is about storytelling. [...] Small introduction, really good guests who are relevant to the story, preferably someone in the story. [...] And keep it short."

Because of that, when choosing a particular angle for covering a story in the forthcoming show, the team was rarely concerned whether the angle was suggested by the underlying data – although, as shown above, the data did occasionally inspire the choice. The primary concern of the production team was who could be a good telephone interviewee and how to approach them. A secondary, yet still important, concern was whether the show presenter and the planned studio guest would be comfortable contributing to the discussion. For example, one week the "Making a Murderer" drama series was trending. The producers did not expect the show's presenter to have prior knowledge of the series, so they decided to invite the shows's producers who would be capable of providing an introduction to the series.

3.3.6.2 Managing timescales

The tight deadlines of the radio production environment required the production workflow to be significantly optimised. In fact, workflow optimisation started already at the data processing stage. For example, the data analyst, by their own admission, had to start the "heavy" initial data processing jobs the very first thing after waking up on Friday and Sunday accordingly. Those jobs included filtering out non-UK tweets, reduction of tweets to their hashtags and collating YouTube and Google Trends. Such a schedule was a compromise in that the data used for chart collation were not as fresh as possible – they only covered the time period until the end of the previous day. However, it guaranteed that the analyst would be able to start data analysis as early as possible in the morning and that therefore the chart would be ready before the production team meeting.

The production team employed its own methods to meet the deadlines. As each producer specialised in a set of topics for which they had a wide contact network, a weekly editor

sometimes had to factor the amount of work one producers could realistically do into their decision on which topics to develop into features. It was not uncommon to hear them say something like "I do not want to create extra work when there is already enough" during the team meetings. Balancing the producers' workload helped with both managing the timescales and securing the diversity of topics discussed on the show. It is worth noting that despite all the efforts made, the producers sometimes had to work on their supposedly free day (Saturday), especially when they aimed to secure guests from other time zones.

The timescales also limited the production team's ability to brief the show's presenter over the content of the topic. It was once observed on a Sunday how the show's editor was briefing the presenter while she was in a taxi on her way to BBC Broadcasting House. One of the interviewed producers claimed that they usually started briefing the presenter on Friday. Given such tight conditions and the variety and unpredictability of topics covered in Hit List, sometimes the given briefing was insufficient. Partly for this reason, the show had to be tightly scripted. The script specified the monologues that the presenter should read to introduce the topic and suggested the questions for the interview, although the presenter was free to adapt those to the flow of the conversation. Moreover, an episode script specified technical details such as when to say the show's catchphrases and when to press buttons to start appropriate pre-recorded segments and jingles. To make sure that the presenter had all the information she needed, some words in the script were even written phonetically. For example, for a story on trichotillomania – a medical condition that creates the urge to pull one's own hair out – the name of the condition was spelled as "Tricho - tillo - mania" in the script. This level of detail was imperative for the production of a show to appropriate journalistic standards.

Finally, the limited time could be a driver behind the very rare – yet still occurring – instances of tweaking the chart. Sometimes a topic that had been in the primary version of a chart on Friday dropped off the chart by Sunday. There was one case when a producer had found a very good interviewee for such a topic on Friday evening. The team had to use their discretion to nonetheless include that topic as the number 40 into the hit list at the expense of the least "newsy" that was meant to be on the chart that week. This was one of the rare cases when the "staying true to the data" ethos was overruled by the purpose of making a good entertaining show, the time limits and the reluctance to waste prior effort.

3.3.7 Summary: Key Lessons

Studying the production of the Hit List show has provided deep insights into the process of social data science in the non-academic context of data journalism. In this setting, the pragmatic considerations often outweigh those of doing "science" in the strict sense of the word. This does

not mean that the findings of this case study are not applicable to the more traditional research settings: similar pragmatic considerations exist in any project, as no one operates with unlimited time and resources. Rather, the Hit List case allows to examine these considerations under a "magnifying lense" of nationwide radio broadcasting. In addition, even when practicalities prevailed, one should not underestimate the amount and complexity of the analytical work that went into the Hit List production and the insights that can be gathered from studying it as well, especially since the experience of the project raises many points on the role of human- and machine data analysis and processing in social data science. The following points briefly summarise the key findings of the case study into lessons for social data science projects:

1. In social data science, algorithmic and human data processing may intertwine in different configurations. While the traditional view of a research process often assumes more human-driven analysis on the initial exploratory steps and more reliance on automated data processing later down the research pipeline, these two stages may come in reverse order – e.g., automated data processing being used to prepare the data for subsequent manual analysis. Moreover, since the results of manual analysis can also inform the automated data processing on subsequent research iterations, effectively a collaborative human-machine data processing loop may emerge. The exact proportions and roles of human- and algorithm involvement should be shaped by project circumstances.

2. Given the notion above, a social data science team should be clear not only about the research questions that they aim to answer, but also about the ultimate goals of their work, which stakeholder needs they aim to satisfy and how this relates to the artefacts of the research work. These parameters are crucial for trading off the priorities for the project such as the comprehensiveness of the research, its rigour, timeliness of delivery and the need to produce results serving the intended purpose for the stakeholders given constraints on available data and research methods. Basic operationalisations such as what is 'newsy' content or what is UK content have to be made with practical purposes in mind as there is no "gold standard" for how they should be made. This is a problem that is not uncommon in the social sciences that basic assumptions and operationalisations are contested (Jarvie, 2011; McIntyre and Rosenberg, 2017).

3. When acquiring data from different sources, a social data science team has to consider that even if the underlying processes that produce the data (e.g. platform users sharing/liking content) and the internal representations of those within the sources are comparable, the sources may return this data to a researcher through different modes of access, thus leading to a variation in the levels to which data are "pre-cooked" – aggregated, filtered and otherwise processed. These differences have to be factored into any comparative analyses or subsequent

data aggregation.

4. The research questions answered in social data projects may not necessarily have a specific "correct" answer. In the Hit List production, such questions may be relatively "innocent" – what stories to consider topical and where to draw boundaries between two topic – but this is not always the case. A social data science team should remember that an absence of a specific correct answer is inherent to such questions – not to the methods of answering them. This notion challenges the false assumption that the answers derived algorithmically are more "objective" than the ones derived through human judgement and thus frees the team from a perceived need to develop purely algorithmic solutions at any cost.

5. If a social data science team wants to develop a machine learning algorithm – for example, text classification – it is crucial for the team members to consider what they would *themselves* rely on when solving the task they plan to delegate to a machine, especially which prior assumptions and knowledge they would employ. This can help not only with choosing the specific details of an algorithm (e.g. feature selection), but also with a "bigger picture" of the algorithm design – e.g. whether a task should be posed as a purely unsupervised one or whether it might have supervised components.

6. If a certain research task gets performed routinely by a certain member (or group of members) within a social data science team, over time some this process may effectively get black-boxed from the rest of the team through forgetting. This forgetting did not produce significant negative consequences for the Hit List production, but in other research settings it might: for example, some team members may misinterpret a pattern observed in data as a substantial finding whereas it is actually a product of prior data manipulations performed by other team members. From the Hit List experience, having whole-team meetings and team members who work closely with different sub-teams helps with overcoming such forgetting.

7. In interdisciplinary teams, technical specialists often systematically work on processing the data while the substantive field specialists only employ high-level data representations (be that for substantive interpretation, journalistic treatment or other purposes). This may disconnect the field specialists from the data and hinder their trust in the data. To restore this trust – and also to facilitate meaningful critique of the data – the technical specialists should provide the field specialists with data overviews that allow "zooming into" particular bits of evidence and moving to lower layers of abstraction when required. The Hit List experience shows the effectiveness of a multi-tab spreadsheet interface for these needs, but the exact choice is likely to be project-dependent.

8. The Hit List production team's ethos of staying true to the data when collating the list of stories for a weekly show and then relying on their own journalistic experience rather than the content of online discussions to determine an angle to cover each story from illustrates that the final artefacts of a social data science endeavour may transcend the pure research outputs and may be informed both by the undertaken research and through other means. A social data science team has to decide how exactly their artefacts should incorporate the research findings.

The experience of the Hit List production case study, in combination with that of the Shakespeare Lives evaluation project, provides the bulk of the evidence used to inform the project management tool for social data science. However, there are still selected project aspects – most notably, that of ethical and legal compliance – for which additional observations are valuable. The following section briefly discusses two other social data science projects that were studied for this thesis.

3.4 Additional Fieldwork Evidence

This short section discusses the social data science projects that provide evidence for some specific aspects of the designed project management tool rather than systematically for the tool as a whole. As a consequence, they are given a brief coverage. The first one of them – the evaluation project for the InfoMigrants initiative – is very similar in nature and in the work involved to the Shakespeare Lives evaluation project (see Sections 3.2). A full discussion of the project would thus be a bit redundant, while some specific observations are of high relevance. The second one – the study of resilience of Dark Net hosted cybercriminal marketplaces – is only discussed briefly here because the primary role of this project was not to inform the designed project management tool but to evaluate it. A more detailed discussion of the project can be found in Section 5.2.

3.4.1 Evaluation of InfoMigrants Initiative

As mentioned in the Shakespeare Lives case study (see Section 3.2), the research project's team leader has a long track record of collaborations with cultural- and media organisations. Running evaluation projects like the one our team performed for the British Council is business-as-usual for her. Therefore, it would not be surprising for a reader that soon after completing the work on Shakespeare Lives, she started to assemble the team for a new evaluation project – this time, for InfoMigrants: a multi-platform news and information resource for those seek asylum in the European Union or are in the process of migration to the EU (Gillespie, 2017). InfoMigrants is a collective effort of three international media organisations – France MÃl'dias Monde, Deutsche Welle and ANSA – that delivers its content through a multi-language website and a number of public Facebook pages, Twitter accounts and WhatsApp channels.

The circumstances of the InfoMigrants case mirrored that of Shakespeare Lives in many respects. Not only did the project have similar objectives (such as the Cultural Value assessment, see Section 3.2.1.2) and employed some of the same data sources (e.g. Facebook public pages, see Section 3.2.4.1) and analysis methods (such as content analysis and online ethnography, see Section 3.2.5) – the teams of the two projects also had four people in common (aside from the PI and myself, another investigator and the lead qualitative analyst). The relationships with the key stakeholder (InfoMigrants being simultaneously a funder and a subject of the evaluation, see Section 3.2.3) were the same. Due to this overlap, many observations made during the two project overlap. Such purely congruent observations are not discussed in this section, although some of them are mentioned directly when discussing the contents of the designed project management tool (see Section 4.3). However, the InfoMigrants evaluation case study did bring some qualitatively new insights that came from the *"benchmarking excercise"* (as our team got

to call this process), which is discussed below.

3.4.1.1 Benchmarking exercise: introduction

As outlined in Section 3.2.5.3, one of the greatest challenges in the Shakespeare Lives evaluation project was lack of benchmarks – i.e. other similarly scoped international cultural programmes – to compare the programme's performance with. Therefore, when assessing the level of the programme's success, the project team had to construct baseline expectations rather than infer them from the data. While the team succeed in doing that, additional assumptions had to be made and the analysis required careful comparison of performances of the programme's own events across different countries and events, which was often non-trivial. Overall, the lack of benchmarks was something that arguably everyone on the team would have preferred to avoid if there had been a chance.

The issue of benchmarking arose again in the InfoMigrants evaluation project, but this time there was more that we could do. To begin with, the representatives of InfoMigrants provided us (as well as the funding organisations) with target values for several key quantitative indicators of online success such as the number of website visits and reach on social media platforms (cf. Stephen et al., 2015). For most modes of quantitative analysis, our team used exactly those figures as the benchmarks. However, in an attempt to further increase the rigour of our research, we decided to complement our core analysis with some form of external comparison of the InfoMigrants' performance – even though, given the modest resources of the project, that had to be done on a limited scale. Conducting this piece of research raised several interesting issues, especially in relation to research design and data acquisition.

3.4.1.2 Designing the exercise

From the research design perspective (cf. De Vaus, 2001), the benchmarking exercise required our team to obtain *comparable data* on online performance of *comparable entities*, i.e. other initiative groups, organisations and campaigns that provided information support to asylum seekers and migrants from economically disadvantaged territories and that operated in the same languages as InfoMigrants. As some of the analysts of the team had prior knowledge of the field, identifying a long list of candidate peers for InfoMigrants was relatively straight-forward. However, upon closer examination, all of them differed from InfoMigrants in one or more fundamental ways:

- Many peers targeted narrower groups of migrants and asylum seekers (e.g. those in refugee camps specifically in Greece);

- Many peers offered information to migrants as part (and often as a by-product) of other services (e.g. help in getting asylum);
- Some peers did not only have offline presence in addition to the online one (which InfoMigrants did not), but specialised on offline activity (e.g. on volunteering in the refugee camps);
- Some peers, who, like InfoMigrants, distributed information in several languages, did not keep their presence in different languages separate (e.g. InfoMigrants had a different Facebook page for each language, while others would post in multiple languages).

This dimensions of variability could influence the differences in online performance of InfoMigrants and its peers. For example, one may expect a long, sustained engagement (and thus repeated interactions) with an organisation that offers direct help to specific migrants and refugees that would come from those who receive the help. This is quite different compared to the sporadic engagement often observed in the media sector. Our team dealt with these discrepancies in two ways. First, using formal (a non-zero intersection in the proposed mission of a candidate benchmark peer and that of InfoMigrants) and informal (expert judgement of the team members) criteria, we shortlisted a fraction of InfoMigrants' peers for whom the online performance data were to be analysed. Second, for each short-listed organisation we provided our stakeholders with a summary that highlighted their differences with InfoMigrants and included appropriate caveats while reporting on our analysis, thus accounting for potential sources of performance discrepancies.

3.4.1.3 Data availability

Our proximity to InfoMigrants provided us with elevated access to their online performance data – some of those data could be obtained from the administrators of their website and of their social media profiles, while others – from their marketing team who used commercial analytical software focused on internal evaluations. While these data were of great help for the bulk of our team's research, most of those data were of little value for the benchmarking exercise since comparable data on the InfoMigrants' peers could not be obtained with this tool set. Our team could only rely on publicly available benchmark data. For example, this implied discarding such a crucial measure of InfoMigrant's success as website visits, since only crude estimates of that metric such as an approximate Alexa Ranking[17] without elevated access to some proprietary data.

[17] For example, see `https://www.alexa.com/siteinfo/infomigrants.net`

After much consideration, it was decided to focus the benchmarking exercise specifically on the Facebook engagement. With the exception of the website itself, the Facebook presence was the one into which InfoMigrants invested by far the most efforts and resources. Moreover, specifically with Facebook, our team was confident in our ability to estimate one of the key success metrics from publicly available data: engagement (cf. Stephen et al., 2015) of individual posts on the Facebook public pages – i.e. a number of interactions Facebook users make with a post. We could operationalise this metric as a summation of reactions, comments and shares that a post had received – something observable through Facebook's user interface and accessible programmatically through the Facebook's Graph API – the publicly available programming interface for reading and writing Facebook Data [18].

It is worth mentioning that our initial plan was to only use the publicly available routes to access the benchmark pages' data, since a more convenient (and allegedly at least as complete) access to the post engagement data for InfoMigrants was available with Facebook Insights – a set of monitoring tools that Facebook provided to administrators of its public pages (cf. Spiliopoulou et al., 2014). However, the Insights data appeared to be inconsistent with those available publicly, with the interaction metrics received through Facebook Insights being systematically higher. To the best of our knowledge, Facebook had never documented the definitive source of this disparity. However, our sporadic observations suggested that the engagement metrics obtainable through Facebook Insights were strictly non-decreasing and thus arguably represented all interactions a post had ever received, while the publicly available ones represented only the currently active post engagement – i.e. these metrics would decrease if someone, say, removed their like or a comment. Thus, for a fair comparison, it was instrumental to treat the InfoMigrants' own public pages the same way as the benchmark sample and to acquire the respective data using the same methods.

3.4.1.4 Data acquisition

Given the set-up of the benchmarking exercise, it was a straightforward decision to develop a web application that would connect to the Facebook Graph API to acquire the data. Compared to scraping, it was more reliable and returned data in a conveniently structured JSON format that did not require much post-processing. The practice of using the Graph API for research had also been well-established over the years of the API's existence. There were only two practical limitations. First, the same data were available to a developer through the API and through the user interface – so, for example, no data from users who had hidden their profiles from non-friends could be accessed. Second, judging from my prior experience, after a number of

[18] https://developers.facebook.com/docs/graph-api/overview

consecutive requests the Graph API tended to start delaying its response. Therefore, the time required to obtain a dataset grew faster than the dataset's size[19]. Both limitations were acceptable for the benchmarking purposes, since only a limited number of *organisational public pages* were studied.

Our data collection application successfully acquired data through the Graph API for the first round of benchmarking exercise in late 2017. However, in spring 2018, when the application was used again, the Graph API only returned error messages instead of data. As it turned out, on March 21, 2018 Facebook put access to the platform's data via the API on hold, unless the requested data belonged to a Facebook user who had used the data collecting web application in the past three months (Xu, 2018; Facebook, 2018a). This condition arose from the "default" scenario that Facebook implied for the use of the Graph API: i.e. developing applications that allowed third party web-services to leverage their user interactions by connecting to a user's Facebook profile. Examples of such enhancements might include providing a "sign in with Facebook" functionality or providing a customer with promotional deals in exchange of posting something on their Facebook timeline. All such interactions require explicit interactions between a Facebook user and the third-party web-app. Thus, the Facebook's decision effectively rendered research-purposed data collection through the Graph API temporarily impossible[20] – unless the research was done privately by an entity that had active app users. Facebook's decision was motivated by the then-recent scandal around the Cambridge Analytica data leakage and seemed quite surprising to our team. Indeed, even if restricting API access to the individual users' data could be considered consequential to the scandal, this was at least questionable for the organisational public pages' data, especially given that such pages often explicitly sought wider exposure and access to their content.

While this posed difficulties for our research it also taught us an important lesson about data acquisition: one should be not operate under an implicit assumption that a publicly available method of data collection would remain accessible indefinitely in the future. Relaxing this assumption means approaching data acquisition strategically, planning the data needs ahead and trying to collect data as soon as they get available rather than when the research actually calls for them.

Overall, the InfoMigrants evaluation project improved our team's understanding of the issues related to designing social data science studies, dealing with inconsistencies of different

[19] To the best of my knowledge, this was never officially documented or tested.

[20] In May 2018 Facebook announced launch of a new version of the Graph API that required web application to through a review process in order to access Facebook data; it is worth noting that no specific procedure for reviewing academic apps was provided (Papamiltiadis, 2018). Later Krishna (2018) reported that Facebook had put additional restrictions on data access again.

representations of the same data and planning data acquisition routines, all of which is valuable for designing the project management tool for social data science.

3.4.2 Study of the Criminal Marketplaces on Dark Net

The final social data science project that I observed and participated in as part of the work on this thesis focused on studying the socio-economic structures of cyber-criminal marketplaces – online platforms that facilitated trade in illicit goods (hereinafter I will refer to it as *the Dark Net project* or *the Dark Net study*).

The case study was conducted when the vast majority of the project management tool for social data science had already been informed and completed. Hence, its primary role was not in informing the project management tool but in providing *formative evaluation of the tool's first version*. These two roles were kept as separate as possible. The evaluation was performed through a series of project management sessions where the project investigators and I used the designed tool to track progress and plan further activities. During those session, we only superficially discussed those project aspects for which prior fieldwork evidence was lacking. By contrast, to inform the missing bits of the project management tool, I used observations of routine work on the project and of those team meetings that did *not* involve the use of the tool – specifically of the meeting dedicated to the issues of ethics and compliance, as those were the least well-informed issues covered by the project management tool at that moment. In inevitable rare cases when one exercise did affect the other, appropriate cross-references are provided.

The evaluation is discussed in Section 5.2 after the project management tool and its content are properly introduced (see Chapter 4). That section also provides a more thorough introduction to the project's background and setup, while the discussion below touches specifically on the primary evidence collected to inform the tool.

3.4.2.1 Ethics and compliance: introduction

Of all the studied research project, the Dark Net case study dealt with the the most sensitive subject matter: we were planning to study evidence of criminal activities. Hence, it is hardly surprising that this project raised the largest number of compliance considerations. The following discussion touches on many of them and shows how they postulated on the levels of both strategic considerations that were almost philosophical in nature and concrete practices that had to be adopted.

3.4.2.2 Stakeholder interests and ethics of cyber-criminal marketplace research

Most typical ethical problems in social research are related to securing interests- and managing risks of the research subjects and participants – these are the main topics covered in standard ethics textbooks (cf. Diener and Crandall, 1978). In studies of criminal communities, this implies considering *the interests of criminal groups*. This may seem to be a futile endeavour since criminals are arguably expected to oppose any attempts at research into their activities. However, in reality there are criminal communities who tolerate and even foster such research. For example, the BlueLight online forum for users of drugs, including the illegal ones[21], has a "Drug Studies" subforum[22]. This subforum is a platform for the researchers in the field to disseminate their findings and to recruit participants for further studies.

In the case of criminal *marketplaces*, it is arguably more difficult to design research that would be tolerable for platform participants since distribution of illicit goods is usually considered to be a more serious crime than possession or use. Furthermore, various groups of platform participants have different reasons to participate which can lead to different attitudes towards research (Martin and Christin, 2016). That said, some of the previous research into criminal marketplaces performed by the Dark Net project's PI, Dr Angus Bancroft (cf. Bancroft and Reid, 2016), was hardly harmful for the markets. It focused on the ways market participants engaged in discussions on the associated forum of one of such marketplaces and did not attempt any analysis that could contribute to revealing the market participants' identities or to anyhow else disrupt their operations. In line with the practice suggested by Barratt (2011), Dr Bancroft even attempted to contact the market administration for a permission to collect data – and even though he did not get any response, his attempt shows how firmly he believed in acceptance of his study by the illicit trade community.

That being said, the Dark Net project was funded by an ESRC Impact Grant and its declared impact component was in establishing collaboration with a law enforcement body. Although in practice communication with this law enforcement body was minimal, it was still recognised as the project's key stakeholder. Balancing between satisfying the key stakeholder's interests and staying ethical towards the market actors thus became a major challenge for the project.

Strategically, the challenge was resolved through the recognition that striving for *overall harm reduction* was something that law enforcement *and* market participants might all agree on, even if the operationalisation of this notion might differ between these two broad groups – and, actually, within the groups as well. This recognition and the according shift in the research focus was largely a product of using the Scorecard Deck and is thus discussed later in the thesis (see Section

[21] http://www.bluelight.org/

[22] http://www.bluelight.org/vb/forums/180-Drug-Studies

5.2.1). However, our developed stance can be summarised as follows:

- Our project is a response to the low efficiency of usual law enforcement interventions in a form of either shutting down a particular cyber-criminal market place or capturing and incarceration of significant market players (administrators or big vendors). Such interventions are focused on disrupting the online trade in illicit goods, but manage to do so only for a very short term and do not lead to harm reduction.

- While the shift of illegal trade to the online sphere may have negative consequences (making the illicit goods more available than before), it may also be a vehicle for harm reduction by reducing street violence and providing customers with fuller information about the purchased goods.

- Ergo, our research should help the law enforcement- and other professional bodies with designing and executing alternative types of interventions that would attempt to steer the trade in illicit goods into a less harmful direction. Such interventions may be disruptive for specific criminal activities (such as selling impure or otherwise overly dangerous drugs), but not necessarily for ordinary market participants such as buyers of drugs for personal use.

One decision that we had to make was concerned with dealing with data that revealed identities of market participants. In principle, such data were not expected to be found. Judging by prior evidence found by Bancroft and Reid (2016), criminal market actors tend to be technologically and otherwise savvy and take the required precautions to retain their anonymity. However, there were known cases of such mistakes – in fact, one of them had led to identification and subsequent capture of Alexandre Cazes, the creator of one of then-biggest cyber-criminal marketplaces AlphaBay (Bellemare, 2017).

Given the overarching ethical stance of our team, it was easy for us to decide to discard and immediately delete any data that were known to potentially disclose the participant identities – unless those were of the types that we would be obliged by law to report on. What seems interesting though is that, in my prior experience of *clear net* research (be that a study of social media- or any other user-generated data), I had never been involved in making a decision on this matter – even though nothing guarantees an absence of unexpected disclosures in the clear net data. The issue of such disclosures is well-known in traditional social research – Guillemin and Gillam (2004) even argue that it is the primary source of ethical dilemmas. Yet, in modern research branches this issue still seems under-appreciated. This highlights the importance of considering the existing traditions of ethics thought in wider disciplinary context when doing social data science.

Another aspect that we had to decide upon was that of research consent. We rejected the idea of contacting marketplace actors or administration for permissions to scrape their data, as, given the involvement of law enforcement, we did not feel that we could actually create trust relationship between us and the market – not to mention that it relieved us of potential legal concerns associated with contacting criminals. Besides, Martin and Christin (2016) argue that a consent for scraping marketplace data is meaningful only if it is received from *all* actors whose data are scraped. This renders recieving such consent practically impossible. In the absence of participant consent, we also decided to only scrape data from the publicly accessible marketplaces – i.e. the ones that did not require invitations to join. While we could attempt to obtain such an invitation, we recognised that participants of closed markets may have reasonable expectations of privacy (Crowther, 2012). Thus, scraping data from such markets without consent would violate our internal ethical standards. By contrast, with publicly accessible marketplaces, it seems fair to assume that the market actors would anticipate scraping attempts to be made at least by law enforcement – compared to that, any academic data collection should be of marginal concern to them.

3.4.2.3 Management of researcher risks

Given the nature of Dark Net research, our team also had to develop procedures to manage the risk for the researchers. Indeed, the discussion above shows that the cyber-criminal marketplace administration should anticipate attempts of data scraping from actors with competing interests – first and foremost, law enforcement bodies, but also maybe competing marketplaces. Therefore, there is no reason to assume that markets would not try to defend themselves against scrapers, including in the form of counterattacks. The most straightforward of such "counterattackes" would be to attempt making the scraping machines executing malicious code to disrupt the scraping process. Even worse would be something harmful for the technological infrastructure of the project as a whole – e.g. something disruptive on a level of the organisational network. In the worst case scenario, the malicious code may help the marketplace to reveal the identities of the scrapers and provide them with means for personal retaliation. This implies the importance of protecting the scraping machine against being compromised – but also of protection of the rest of the project infrastructure and of the researchers' identities in the case the scraping machine *does* get compromised.

To respond to such risks, we used several layers of protection. As an initial layer of protection, we accessed the scraped web-pages through a Tor browser with JavaScript turned off, so that the pages we visited could not trigger malicious client-side code to be executed. This is a fairly standard practice on Dark Net; from our research experience, most of the marketplaces do not have dynamic content to start with specifically to make the websites usable without

JavaScript. As an additional layer of protection, we developed and executed our scraping code on an anonymous remote machine hosted in the Google Cloud. We adhered to a policy not to store any personal identifiers there. For example, we never actually pushed changes to the scraping code directly from the Google Cloud machine to a BitBucket repository that we used for version control and code management – rather, we first triggered a secure copy of the code on a server at the University of St Andrews from the Google Cloud machine to that server. To further protect outselves, we stored all the data on the same server and analysed them remotely.

It is interesting to observe that, while anonymising connections through Tor was necessary from the start as the marketplaces were not accessible outside of the Tor network, we also came to rely on it for managing our own risk as researchers. We extended this even to scraping clear net websites such as Deep Dot Web[23] – a portal that posts news about illicit goods and their trade, lists hyperlinks to the Dark Net marketplaces and provides a discussion space for those. As Deep Dot Web by itself is not a platform for conducting criminal activity – only for discussing it – it is made accessible outside the Tor Network. Still, as our risk management practices prescribed hiding our identities from the administrators of scraped websites, we still used the Tor network for this – even though using it slowed down the connection.

Another risk that we had to protect ourselves against was accidentally scraping data that would by themselves be illegal. While this risk, as much as the risk of accidental disclosures discussed above, could not be eliminated, we did take some measures to minimise it. For example, we specifically did not scrape any visual content since that could depict illegal scenes. Additionally, we did not scrape marketplaces or forums that dealt with specifically sensitive product categories such as weapons.

A final point that has to be made is that our identities would still have to be revealed should we ever publish our work (or about our work, as is done in this thesis). What allowed us to feel comfortable with this risk was, again, the overall ethical stance of our team and a lack of desire to fundamentally disrupt criminal marketplace operations. In our assessment, our intent would not be evident from the process of data collection, but would be from the published results. Additionally, we considered the experience of previous research that had not shown cyber-criminal marketplace actors to be unhappy with research into their activities (Bancroft and Reid, 2016). Yet, this example reminds us that some risks cannot even be significantly minised – it is only possible to account for them and to make an informed decision on whether to accept them or not.

[23] `https://www.deepdotweb.com`

3.4.2.4 Institutional ramifications

While the discussion above focuses on the substantive concerns that our team had to address, we also had to comply with the formal policies and procedures of our universities. As shown below, those proved to add a layer of complexity and gave an impression that the UK institutions are still ill-equipped at regulating conduct in social data science.

Any ethical review at the University of St Andrews starts with filling it an Ethics Application Form. This form is standardised across all academic schools and types of research projects with an exception of the projects that pose risks not only for humans (e.g. the projects that involve animal testing). Applying such a standardised form to the Dark Net project (and arguably many other social data science projects) was sometimes problematic. For example, most of the questions were formulated in such terms as "project participants", "locations of research" and others that were not really applicable to mass-scale online data collection, but to traditional research. As a result, we decided to leave many questions without an answer and to provide a several-page long free-form ethical statement instead. While the ethical complexity of the Dark Net project would oblige us to provide us with *some* free-form statement anyway, it could arguably be cut in short quite significantly if there was a better-suited form to begin with – not to mention that the decision to leave questions without a response also came with hesitation and thus took some time to be made.

Another instance when the institutional policies appeared to be ill-equipped for our project revealed itself when, after the initial ethics review, our team was recommended to run the data acquisition code on a specific university's web server that was delegated for Dark Net research. This suggestion went against our ideas of protecting the project- (and the wider university-) infrastructure by acquiring data on a virtual Google Cloud machine. As it appeared, the university's policy had been tailored for a particular type of Dark Net research – terrorism studies. The usual methods of such research do not include massive data collection and thus do not cause a high risk of counterattacks from the studied Dark Net platforms. On the other hand, the data that are acquired in such studies are often illegal to be in possession of – thus, by obliging its staff to use a specific university machine for terrorism data collection, the University of St Andrews makes sure that the researchers, in the case of suspicions from the law enforcement, have a chance to make the case of collecting the data for professional use. As can be seen, the circumstances of our study were significantly different to that in terrorism research. Luckily, the university proved to be sufficiently flexible to recognise this difference in circumstances and allow us to adhere to the plan. However, if we had known about this university policy in advance, we could have included this bit into our ethical statement from the start and thus potentially get an approval earlier.

3.4.2.5 Learning about compliance

As a brief final piece of commentary, the compliance issues raised by this project were not only multidimensional – for our team, many of them were novel. Of us all, only the principal investigator had substantive prior experience in researching the Dark Net (cf. Bancroft and Reid, 2016). Moreover, his prior research in this area only employed traditional sociological methods that could not raise the issues associated with large-scale data acquisition and were better governed by the existing traditions of research ethics overall. The generic social data science experience that the computer scientists on the team had, as well as their prior limited work on compliance (cf. "Ethics and legal compliance" in Voss et al., 2016), could partially compensate for that. However, there were still questions of the possible interactions between these two sources of compliance issues: our team lacked a framework to think about the compliance systematically and to communicate our stance to the hosting universities. A remedy came in the form of the paper by Martin and Christin (2016) on ethics in cryptomarket research. It discussed four dimensions of the Dark Net research ethics thus helping us to shape our thinking on ethics and the corresponding ethical statetemnt into a self-consistent and comprehensive form. As such, it helped us cope with our "known unknowns" in ethics of crypromarket research (i.e. the issues we understood we had but were not entirely sure how to cope with). It also exposed us to our "unknown unknowns". For example, prior to reading the paper we had not quite recognised that different actors active on cyber-criminal marketplaces may have different attitudes towards research and ethics that had to be taken into consideration.

Two points can be carried away from this. First, given the complexity and diversity of social data science, it may be valuable to do literature review and wider context survey specifically of compliance issues in the field of undertaken research. While this idea may sound self-evident, it is actually not uncommon to see prior review being focused exclusively on the already obtained findings and on the employed methods, but not on compliance. More broadly, the discussion above shows that the understanding of what the ethical issues are in a social data science project and what the appropriate and practical responses are will often be evolving. Social data science is a young discipline of extreme diversity and thus there cannot be a ready-made template for identifying and dealing with ethical issues. This implies that *social data science teams* should do *ethics* as continuous reflection on ever-evolving research activities. This also suggests that *insititutions* should support equally iterative *ethical review processes* for social data science that do not assume full up-front ethical clearance. The lack of such processes is a practical problem that social data scientists are forced to cope with.

Plus, the context survey on compliance should not necessarily be limited to the same corpus of work a research team would get acquainted with anyway – it may involve various forms

of guidance and policy documents. Second, while this is not an observation directly about compliance, it is interesting to note that what Martin and Christin's paper has done for our team is similar to what the designed project management tool aims to do for social data science projects as a whole. The helpfulness of this paper further motivates the relevance of the tool.

3.5 Chapter Postface

The brief account of the additional evidence from the InfoMigrants evaluation project and the Dark Net study concludes the discussion of the fieldwork undertaken to inform the project management tool for social data science. To reflect on the analysis above, I would like to highlight that the core recurrent issues threading through the case studies appear to have a dual nature. On the one hand, they come from the practical complexities of doing social research with the new methods and the new forms of data. This is particularly true for the issues associated with accessing the data – be that the limitations that the platforms pose on the publicly available data acquisition modes or even the mismatch with the implied scenarios of use for the commercial data access software. However, in many cases the discovered issues could be tracked down to the fundamental problems in studying the social – the same problems that pose the questions of how much the social studies are actually a science (and by what criteria, cf. Jarvie, 2011) and raise the tension between the desire to cast them as a such and the liberties that may come from abandoning the notion of scientific nature (Wilde, 2010).

Large datasets and the alleged rigour they allow for cannot (even in principle) resolve the need for making decisions that reflect a researcher's (or a team's) stances and beliefs rather than being informed by a scientific process. To the contrary, the new forms of data motivate a whole new round of such decisions: should a "like" be weighted lighter than a comment? What is a "UK" tweet? Should the audiences of a soft power organisation be expected to explicitly praise the organisation's country in their naturally occurring interactions? What are the valid criteria of success on social media? Is the knowledge exchange by drug users facilitated by the Dark Net a good or a bad thing?

From this perspective, the term "social data science" defended early in the thesis (see Section 1.1) may also be considered a bit of a misnomer – but a useful one, as its "data science" part is a reminder of where the new analysis modes come from – it is just that the methodological and philosophical implications that come with the "social" bit should not be disregarded. And accordingly, the designed project management tool should be both a reminder of those implications (by explicit representation of the underlying principles of a social data science project and of their operationalisation) and the mechanism of navigation through them (by

giving members of a social data science team a common artefact to structure their discussion of those issues around and to keep reminding themselves of the agreed stances). The next chapter introduces the tool and explores its contents.

CHAPTER FOUR

SOCIAL DATA SCIENCE SCORECARD DECK: THE PROJECT MANAGEMENT TOOL

The previous chapter has discussed several social data science projects that I studied as a participant observer. Now I will use the experience of these projects and the lessons drawn from them, as well as the review of relevant literature presented in Chapter 2, to inform the Social Data Science Scorecard Deck, a project management tool that adapts the SEMAT Essence model (OMG, 2015; Jacobson et al., 2013) to social data science. Section 4.1 provides an overview of the Scorecard Deck by exploring the areas of concern and the core project aspects (*alphas*) of a social data science project that are covered by the Scorecard Deck (4.1). Then I discuss the process of using the tool, the contexts in which it might be done and how this is affected by its implementation as a Google Spreadsheet (4.2). In section 4.3, I go through each alpha, discussing its respective scorecards and relating their content to the fieldwork and, where relevant, to the literature.

It is worth noting that while the chapter presents the finalised design and contents of the Scorecard Deck, in practice its development was iterative. Most importantly, some specifics of the scorecards were designed only after the first round of evaluation discussed in Section 5.2. Such cases will be clearly indicated.

4.1 Essence of Social Data Science

I would like to start the discussion of the Scorecard Deck (or simply the 'Deck' from now on) in a manner that parallels the SEMAT approach – i.e. by providing a conceptual overview model that specifies the key areas of concerns in a social data science project (for discussion of the corresponding SEMAT model, see Section 2.4.2). For each area of concern, this model lists its core aspects (i.e. *alphas*) and depicts the relationships between the alphas and areas. Following SEMAT's tradition, the conceptual model is called the *Essence of Social Data Science*. It is represented in Figure 4.1.

Out of the five represented areas of concern, the "Demand", "Response" and "Endeavour" areas are in direct correspondence with the three areas of the Essence of Software Engineering. Each of them represents one of three fundamental components of any project in any field – the need that motivates the project, the outcomes of this work and the process of work itself. Two other areas – "Analytics" and "Resources" – are unique to the Essence of Social Data Science model. The following sections will examine the five areas one by one, discussing the respective alphas and their relationships.

4.1.1 Demand Area of Concern

The *"Demand"* area of concern captures the need for running a social data science project. The area contains two alphas – "Research Goals" and "Stakeholders". The area is analogous to the "Customer" area in the SEMAT model. A different name is employed since the notion of a customer is a problematic one and may be misleading. In its everyday meaning, this term is not always directly applicable to a social data science project. The term customer arguably is normally associated with a person or an organisation that purchases a good or a service for their own benefit. In case of project-based activity, this implies being simultaneously a funder and a lead beneficiary. In the non-for-profit sector, including academia, those are often two different roles. While the funder always has a stake in a project's success, it is not necessarily *their* needs that have the biggest impact on the direction of a project. For example, the Dark Net project (see Section 3.4.2) was funded by an impact grant from the Economic and Social Research Council (ESRC), yet our work's primary objectives were focused on satisfying the needs of our a law enforcement body that we planned to collaborate with. Although by succeeding in that we would have also satisfied the expectations of the ESRC, those expectations were almost never explicitly discussed in the project meetings. Given this, "Demand" seems to be a more appropriate generic term.

The "Research Goals" alpha represents what, in principle, needs to be achieved by the research.

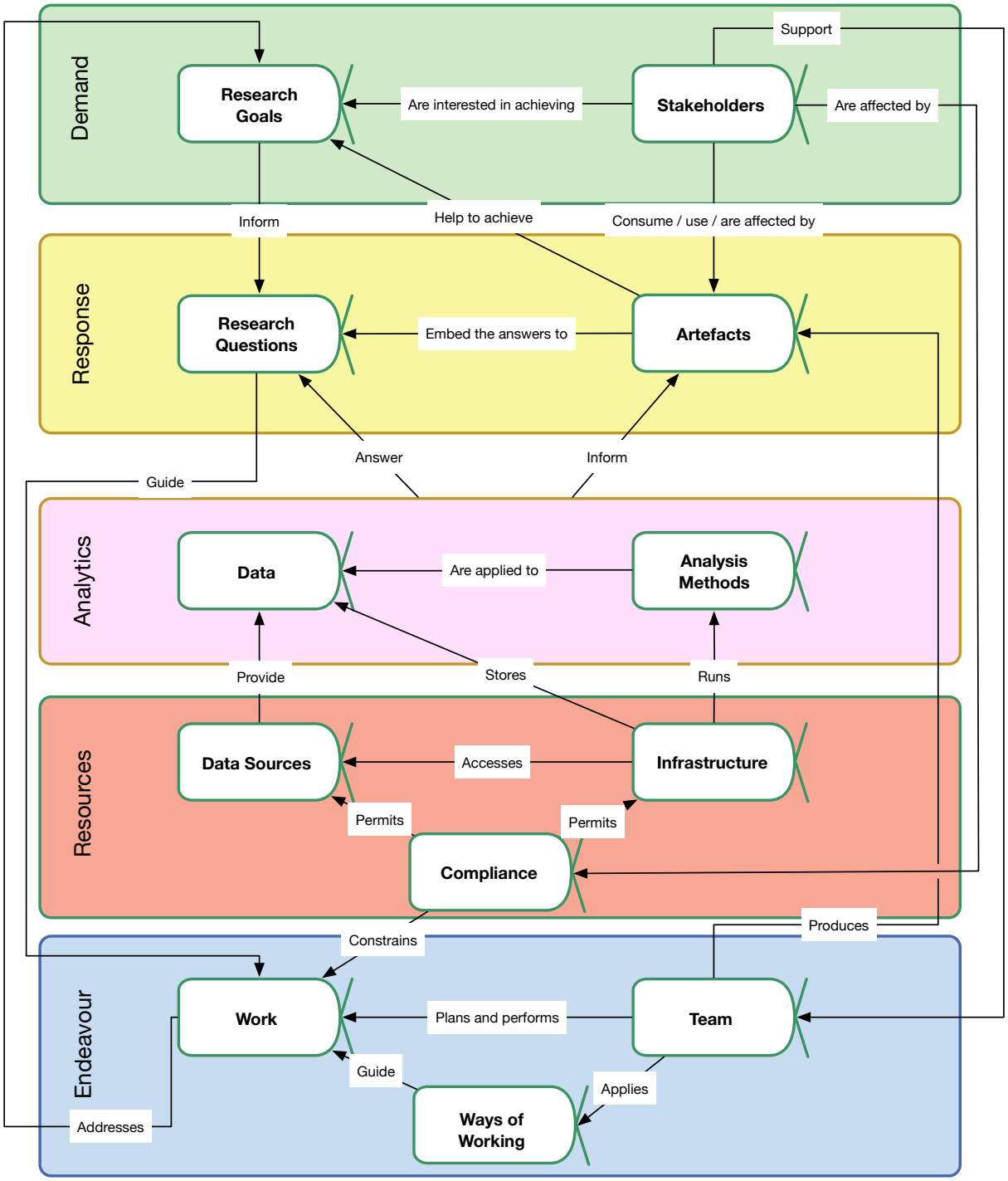


Figure 4.1: Essence of Social Data Science diagram.

It corresponds to the "Opportunity" alpha in the SEMAT Essence model. Research goals motivate the social data science project and provide it with a direction. They justify the project's existence by capturing the needs of the stakeholders and the underlying problem that the project tries to solve. They also help to determine the scope of the project so that its impact is valuable while carrying it out is feasible.

The fieldwork shows that the most high-level goals of a social data science project are not always formulated in research terms and require more than just the scientific process to be achieved. For example, the ultimate goal of the work discussed in the Hit List case study (see Section 3.3) was not to simply find which topics had been the most discussed online in the UK within each week (something that can be straight-forwardly adapted into a research question) – it was to produce a high-quality weekly radio show based on those topics. The "Research Goals" alpha represents both pure research- and overarching project goals, as those are very deeply intertwined and go through similar states.

The *goals inform the questions* that the research aims to answer. Some of the prior work in which I participated before this PhD demonstrates the problems that may arise should the link between the questions and the goals weaken. It is quite illustrative that almost a half of the methodological paper that discussed that work is devoted to its limitations (Dennis et al., 2015). By contrast, all the case studies discussed in this thesis treated their goals with respect – be that the attention to the stakeholder needs in the Shakespeare Lives- and the InfoMigrants evaluation projects (see Section 3.2 and 3.4.1) respectively) or the highly collaborative work towards producing the Hit List show.

The "Stakeholders" alpha represents the potential beneficiaries of a social data science project, the users of the artefacts and, in accordance with Freeman's (1984) ethical strategic outlook on stakeholder engagement, other parties affected by a social data science project and thus *are affected by how compliant with ethical and legal standards* the project is. Stakeholders may *support the project team* through funding, guidance, participation in research activities and other means. It is the *stakeholders who are interested in the achievement of the research goals*, and thus they are the final judges of their completeness. The stakeholders also *consume research artefacts*, which can, for example, explain the increase in the frequency of stakeholder meetings towards the end of the project observed in the Shakespeare Lives case study. Even when the pure research component of a social data science project is complete, the project can still go through iterations to make the final product – in that particular case, a website presenting the research outcomes – as useful and digestible for the stakeholders as possible.

4.1.2 Response Area of Concern

The *"Response"* area of concern captures the attempt to fulfil the need in a social data science project. The area contains two alphas – "Research Questions" and "Artefacts". The area corresponds to the "Solution" area of concern in the SEMAT Essence model. The change of term to an (again) broader one is driven by the realisation that in social data science, as much as as in science in general, it is not always possible to arrive at a complete solution to the posed problem. For example, while in the Shakespeare Lives evaluation project our research team managed to go one step beyond the findings and to provide recommendations on running similar cultural programmes to the British Council, by no means can we claim that this list of recommendations is exhaustive and can guarantee the best possible online performance. Moreover, in a more academic research context where social data science projects face even higher level of uncertainty, there is always a risk to arrive at non-findings and essentially not to "solve" the research problem at all.

The "Research Questions" alpha represents the questions that the research aims to answer. This alpha loosely corresponds to the "Requirements" alpha of the SEMAT model. The bulk of the work on a social data science project is done to answer the research questions, so they *guide the work on the project* from start to finish – in a similar manner to the requirements, which guide the process of software engineering. Moreover, as much as the requirements, the research questions are iteratively refined throughout the course of a social data science project. This quality of the research questions is discussed later when introducing the respective scorecards (see Section 4.3.3).

The "Artefacts" alpha represents the outputs of the project that *embed the answers to the research questions*, i.e. the research findings. The complexity of embedding the findings depends on the type of artefacts – this may be more straightforward for the usual research outputs such as academic papers, but gets trickier when the raw findings have to be balanced with other considerations such as the quality of journalistic work, as in the Hit List case study. This notion emphasises that artefacts *help to achieve the project goals*. From this perspective, the "Artefacts" alpha loosely corresponds to the "Software System" alpha in the SEMAT model.

That being said, the differences between the social data science artefacts and software systems in engineering projects are also profound. First, the types of social data science artefacts are numerous: while they may take a form of apps and software modules, especially in an industrial environment, they may also be reports of various forms (report documents, papers), media artefacts, and others. Second, a software system can be metaphorically named "the alpha and omega" of a software engineering process – the vast majority of the work directly feeds into

it from the start of the project (the work iteratively goes from designing and early prototyping to implementing and testing the components of a system). In contrast, in social data science a lot of work goes *not* into the artefacts themselves, but into *informing the artefacts with the project's analytical work*. The artefacts, in turn, are often developed at the end of the research cycle. This was the case in the Shakespeare Lives evaluation, as the work on the final project artefact (evaluation website) commenced only during the last round of project reporting. In the Hit List case study, the artefacts (show's episodes) were released each week; however, the work on each particular episode started only when the list of the most popular topics for its week had been collated (see Section 3.3.1).

4.1.3 Analytics Area of Concern

The *"Analytics"* area of concern captures the substance of the research process that is required to *answer the research questions* and to *inform the artefacts*. This area of concern is unique to the Essence of Social Data Science and does not have a counterpart in the SEMAT model. It could be loosely compared to the "Work" alpha of the SEMAT model, since both, in a sense, represent what has to be done to achieve the goals. However, the Essence of Social Data Science makes a distinction between the logistics aspect of the work involved (planning, task distribution, monitoring, communication, as well as tying up the research activities within a project with other activities that are required to produce the artefacts) and its actual research substance. The former is still represented with the "Work" alpha inherited from the SEMAT model, while the rest is represented by the alphas within the "Analytics" area of concern. This distinction allows to more specifically guide a social data science project through the research process.

A reader may argue that "Analytics" should be represented as an alpha within the "Response" area of concern rather than a separate area of concern by itself. Indeed, the area's loose counterpart in the SEMAT model is an *alpha* rather than an area of concern. Treating "Analytics" as an alpha is, in principle, possible since one may talk about the progress of a project's research process as a whole. However, treating "Analytics" as an area of concern allows to look separately at its core components by treating *those* as alphas. The Essence of Social Data Science distinguishes two such alphas – "Data" and "Analysis Methods". Treating them separately allows for a more detailed guidance of what is arguably the central aspect of a social data science project.

That being said, the "Analytics" area maintains a property that is typically associated with an alpha: as the Essence diagram depicts, this area has direct relationships with other alphas (outgoing to "Research Questions" and "Artefacts", incoming from "Compliance"). These relationships are intuitively clear and are confirmed by fieldwork. For example, in the Hit List case study, both inclusion of new data (i.e. data from new data sources) and changes in the data

aggregation methods tangibly affected the contents of the radio show. In general, a connection in the Essence model that has an area of concern as one of its nodes should be interpreted as if every individual alpha within this area of concern had such a connection – it is just a tool to compress redundant information. This annotation does not perfectly align with the the SEMAT formal language (see Section 2.4.2). However, since the end-goal of this thesis is to produce a managerial tool for social data science projects rather than to formally model such projects, practical utility and ease of representations becomes a higher value than full compliance with the details of the standard.

The "Data" alpha represents all the data that are available for the project. Data may come from both external and internal data sources and may be produced by the project interim work. For example, in the InfoMigrants evaluation case, the data on the initiative's social media performance came from external sources (the platform APIs and proprietary third-party analytical tools), while the data on website visits were available internally to the organisations that carried out the initiative. In the Hit List case, the data analyst produced a chart of popular news topics which in itself became data for further journalistic treatment of the show's producers.

Data are a vital core aspect of a social data science project and often a problematic one. In all case studies examined in this thesis, the teams had to deal with data that were less than ideally tailored to answer the research questions. Not all social media posts studied to produce the Hit List were news-related (see Section 3.3.4); not all tweets that the Shakespeare Lives evaluation team studied to estimate public's emotional reaction towards the cultural programme were of types that could display any sentiment other than neutral (see Section 3.2.5.3); in the InfoMigrants evaluation project, the openly available data on social media engagement that had to be used for benchmarking needs could only approximate to the complete engagement metrics available through Facebook Insights and employed for the other analyses (see Section 3.4.1.3). While there is no universal recipe for how to deal with such problems, they at the very least indicate the kinds of questions that research teams should ask themselves about what data they need, what they can realistically have and how to close the gaps. Those questions are incorporated into the respective scorecards (see Section 4.3.1).

It is worth noting that the "Data" alpha specifically treats research data as one of the components that feed into the analysis performed within a social data science project. In practice, data are also often one of the project artefacts. For example, the UK's Economic and Social Research Council (2018) encourages data sharing and expects its funded researchers to do so unless ethical- or other justifiable circumstances do not allow it. If this is the case in a social data science project managed with the Scorecard Deck, it may be useful to treat the project data as subject of both the "Data"- and "Artefacts"-related scorecards.

The "Analysis Methods" alpha represents the particular analytical procedures and techniques that are *applied to data* to derive knowledge from them. Social data science allows for various modes of analysis and may include but are not limited to qualitative analysis techniques, data visualisation techniques, descriptive statistics, classical inferential statistics and machine learning. The issue of choosing and applying those techniques can be contrasted with the wider issues of research design reflected in the "Data" and "Research Questions" alphas and with the issue of data acquisition modes and methods reflected in the "Data Sources" alpha. Since in practice these issues are deeply intertwined, it presents a challenge to representing the "Analysis Methods" alpha in the Scorecard Deck. More on that challenge – as well as on the reasoning why specifically the *analysis* methods are considered as a separate alpha, rather than research methods in general – can be found in Section 4.3.4.

4.1.4 Resources Area of Concern

The *"Resources"* area of concern captures the resources necessary to run a social data science project. Three alphas belong to it – "Data Sources", "Compliance" and "Infrastructure".

The original SEMAT model does not delegate a separate conceptual entity to project resources (neither an area of concern nor an alpha). The issues of resourcing are instead attributed to the alphas to which the respective resources are related. Thus, the "Opportunity" alpha covers the financial resources, as this alpha stresses the need to evaluate the financial feasibility of exploiting the opportunity. The "Team" alpha cover the human resources required to carry out a project. Finally, the "Software system" alpha represents not only the solution to the opportunity, but also the required infrastructural (software and hardware) resources.

For social data science, reducing resourcing issues to other alphas is not always possible. The aforementioned example of infrastructure – a system of software and hardware components that underpins the research process – is the prime example of that. Infrastructure proves to be a vital resource for all sorts of social data science projects almost in every context. Even the InfoMigrants evaluation project, which relied on very limited use of computational methods, required a complex infrastructure, for example, for data acquisition (see Section 3.4.1.4). The human coding of qualitative data discussed in context of the Shakespeare Lives case study also relied on the use of specific software despite being a labour intensive rather computationally intensive analysis mode (see Section 3.2.6).

Another reason to distinguish resources as a separate area of concern is the specificity of their key types: a social data science projects depends on its data sources – as they feed the project with data – and its technical infrastructure – as it allows to harvest and process these data. Compliance

with ethical, legal and licensing norms and permissions thus becomes a meta-resource that governs the use of the two main resource types and constrains what they may be employed for. On the other hand, non-compliance may, in the worst case, render project artefacts unusable by the stakeholders; from this perspective, compliance may be viewed as a positive, risk mitigating resource. Consequently, "Data Sources", "Infrastructure" and "Compliance" are considered as the three alphas of the "Resources" area of concern.

The "Data Sources" alpha relates to any system that (a) motivates and facilitates the creation and/or storage of data and (b) provides a mode of access to these data that is available to the research team. The data sources for social data science are as varied as the data themselves – for example, they include social media-, search engine-, web marketplace- and other online platforms, mobile and sensor devices, and transaction recording systems. Many observations made within the case studies point at the complexities of working with various data sources. This results in one of the most detailed sets of respective scorecards specifically for this alpha (see Section 4.3.2).

The "Infrastructure" alpha, as already mentioned, is a multi-component system facilitating the research process. Components of the infrastructure are primarily software and hardware. The infrastructure is used to access the data sources, to acquire and store the research data and to execute the research methods. In terms of the actual progression throughout the course of a social data science project, the infrastructure goes through very similar states to those of the Software System alpha in the SEMAT model, as both may be based either on software development or on construction by configuration (Sommerville, 2008). For this reason, the content of the "Infrastructure" scorecard set is almost in full inherited from the SEMAT model.

The "Compliance" alpha represents all the available compliance resources which normally take the form of ethical and legal clearances, permissions and agreements. Compliance governs the legal and ethical acceptability of accessing the data sources, storing the data and subjecting the data to analyses using the selected methods.

The "Compliance" alpha does not have an analogue in the SEMAT model. Moreover, the coverage of compliance-related issues in SEMAT is very limited even within other alphas. This may seem counter-intuitive, since at least legal compliance is undoubtedly relevant to software engineering – the issue of licensing external software components may be critical in determining the cost structure of running a software system and thus in the buy-build decisions (Daneshgar et al., 2013). Since construction by configuration (Sommerville, 2008) is a prominent model in today's software engineering, those decisions play a crucial role.

There are at least three reasons that justify having a separate "Compliance" alpha. First, in social

data science, compliance is not simply ensured once at a certain point in a project when license agreements are read, understood and put into practice. Instead, monitoring and maintaining compliance is a continuous task (see Section 2.3). Second, the interests of stakeholders are often competing (see Section 3.4.2.2). Third, the social data science teams have to be *proactive* in their treatment of compliance questions. That is, they should try to anticipate potential changes (crucially, restrictions) from the compliance side and have a plan of action (see Section 3.4.1.1). As an overarching theme for all three conclusions, it may be very beneficial for all team members to have a strong and well fleshed-out shared understanding of the compliance issues involved in a project and of the team's stance on how to deal with those issues, so having the compliance issues represented in the Deck may be of great help for that.

4.1.5 Endeavour Area of Concern

The last area of concern covered in the Essence of Social Data Science model is "Endeavour". The area is concerned with the practicalities of carrying a social data science project.

This area inherits all of the alphas and the vast majority of properties from the corresponding area of concern in the SEMAT model. The reason for this is that the SEMAT model captures this area in a way that is arguably applicable to all projects, not only software engineering ones. **The "Team" alpha** is in one-to-one relationship with it SEMAT counterpart. It represents the people who are occupied with the project and perform the required work. **The "Work" alpha** represents all the tasks that have to be performed within the project and thus the process of addressing the project goals. **The "Ways of Working" alpha** represents the guiding principles that the team adhere to when deciding how to plan, distribute and perform the work. Given the highly adaptive, iterative nature of doing social data science, I decided to use the word "ways" in plural as compared to "way" in the SEMAT model.

4.2 Using the Scorecard Deck

Like the SEMAT Essence model (see Section 2.4.2), the Social Data Science Scorecard Deck is meant to be compatible with any process model of organising and managing a social data science project. As a tool that captures the progress of a project, it should help social data science teams to see how far they have progressed on different key aspects of their projects and suggest potential further steps without requiring any particular process of implementing those steps.

As a *holistic* project management tool that puts equal consideration on "hard" and "soft" issues involved in social research, the Social Data Science Scorecard Deck supports the principles of social data science responsibility (see Section 2.1.3.1):

- The Deck helps social data science project teams to achieve shared understanding of their work by providing a convenient talking point about the core aspects of a social data science project and their relationships (see Section 5.2.1).

- The Deck supports research transparency by helping social data science teams to structure their external reporting (see Section 5.2.1.3).

- The Deck supports reproducibility by helping research teams to track and record their decisions and by encouraging them to preserve their data, practices and infrastructure (see Tables 4.5, A.35 and A.19).

- The Deck supports iterative and reflexive use of analytical modes through suggesting how data demands, research questions and analysis methods are critically refined (see Sections 4.3.1, 4.3.3, 4.3.4). The Deck's spreadsheet interface (see Section 4.2.2) makes it easy to roll back to earlier project states and iterate through those.

- The Deck supports combating research biases by helping researchers to put together varied portfolios of analytical methods and data sources that allow for triangulation and complementation (see Tables 4.6, 4.7 and 4.17) and by tracking decisions involved in research operationalisations (see Table 4.2).

- The Deck supports social data science teams in assessing performativity through recognition of the stakeholder groups, assessment of their needs and understanding what impact research artefacts might have on them (see Section A.2).

4.2.1 Process and Context of Use

The process of using the Social Data Science Scorecard Deck is almost identical to that of the SEMAT Essence model. Each of the core project aspects (alphas represented in Figure 4.1) is represented by a top-level scorecard that shows a number of states tracking its progress. Each of these states has its own card. Each state card has a number of conditions a project has to satisfy for a state to be achieved.

One innovation compared to the SEMAT model is that states also have ***prerequisites***: the suggested states of *other* alphas that have to be achieved before the assessed state can be considered achieved. Prerequisites interlink the sets of scorecards for different alphas and allow to better judge the state of a project as a whole.

To use the Scorecard Deck, one has to select an alpha whose progress they would like to assess. They would need to consider the conditions and the prerequisites on each of the state cards

comparing them to the project's status. This provides an indication as to where some activity might be required to progress the project. In principle, the conditions can be ticked off as satisfied in any order, but the idea is that for the most part a state can only be reached if all previous states are.

The Social Data Science Scorecard Deck can be used in multiple environments:

- During the project team meetings, the Deck can be used to track the progress of all the critical aspects of a social data science project and to identify the aspects that fall behind the others. It can also be used to plan further activities and their timespans. Furthermore, the scorecards can serve as a map to distribute responsibilities among the individual team members.

- The Deck can be used in the routine work of individual team member. A team member can use the alphas, states and conditions that they are responsible for as a convenient structure for documentation of their work and for their progress reporting. The project manager or a team leader can use the scorecards for continuous assessment of the project state.

- The Deck can be used outside of the scope of a particular project as a way to classify and organise best practices and ways of working that a team or an organisation has developed and to store pointers to relevant guidance.

Depending on the complexity of a project, the Deck can be used in the context of the project as a whole, of a particular branch of the project and of a particular iteration.

4.2.2 Spreadsheet Interface

While the original SEMAT model comes in the form of paper artefacts – actual physical cards – the Social Data Science Scorecard Deck is presented as an interative Google Spreadsheet. While tangible cards do have an advantage in that it is possible to use them in engaging interactive activities within the project meetings, in the case of social data science those are overshadowed by the advantages that the electronic form brings:

- *Opportunity for remote collaboration.* From the experience of the case studies, social data science teams are often geographically disperse. This is arguably especially common in the academic context, with team members from different universities working together. As Google Spreadsheet allows concurrent collaboration on the same document, it becomes possible to interactively use the Scorecard Deck in remote team meetings.

- *Opportunities for revision.* Doing social data science is a highly iterative process. A condition that seems achieved today may cease to be so tomorrow due to previously undiscovered

methodological considerations, adjustments to research goals and questions, newly discovered data sources and so forth. An opportunity to tick on / tick off the same condition as many times as required is thus crucial.

- *Granularity in progress tracking.* A consideration related to the previous one – with electronic artefacts, it is easier to implement intermediate states of achievement for particular conditions. For example, the Social Data Science Scorecard Deck allows to mark a condition to be in progress, achieved subject to further revision or achieved definitely.

- *Opportunities for annotation.* Spreadsheets are very flexible tools. Users of a Social Data Science Scorecard Deck spreadsheet may benefit from this flexibility by using the space near the scorecards to put more detailed notes on their progress, to provide their operationalisations to the assessed states and conditions or to insert hyperlinks to related project files and documentations. In the latter case the Scorecard Deck effectively becomes a table of contents for the research project.

- *Opportunities to adjust the scorecards.* If a project team feels that a certain component of their project is either misrepresented on not represented at all in the existing scorecards, they may always make changes to the tool itself.

- *Automation.* Since the Social Data Science Scorecard Deck introduces the concept of prerequisites that tie together scorecards for different alphas, there is scope for automating navigation, representation of progress and cross-referencing in general.

- *Ease of export.* The assessment of conditions, as well as the supporting notes, can be easily exported from a spreadsheet, for example to be included into project reporting.

- *Sharing.* If a new member joins a social data science project team, it is easy to share the project's scorecards with them. This can be a useful measure to introduce the newcomer to the current project state.

In the spreadsheet representation, each tab is dedicated to one alpha. The alpha card is at the top of the sheet, with the state cards stacked vertically beneath it. Figure 4.2 presents an example alpha ("Research Questions") and state ("Research Questions 1/5 : Outlined") scorecard from the Google Spreadsheet representation. In this imaginary example, a user has worked through the state scorecard. For each condition, they assessed its accomplishment from a predefined set of options *(No, In progress, Yes (to be revised), Yes (definitively), N/A)*. Similarly, for the prerequisites (two bottom rows), they chose whether to consider or discard those in assessment of

RESEARCH QUESTIONS	*All questions that a research project aims to answer. Research questions guide the work on a social data science project.*	**7%**
1. Outlined	The research questions are formulated as a broad framework that sets the direction for the project.	**30%**
2. Refined	The questions are redefined and specified to an operational level.	**1%**
3. Matched	The refined research questions are re-validated to match the research circumstances.	**0%**
4. Answered	Through applying the analytical methods to the acquired- and quality assured data, actionable answers to the research questions are derived.	**2%**
5. Utilised	Answers to the research questions are in use.	**0%**
Research Questions 1 / 5		
OUTLINED	*The research questions are formulated as a broad framework that sets the direction for the project.*	**30%**
- Demand understood	The research needs and goals can be translated into specific knowledge demands.	**Yes (to be revised)**
- Scope of existing knowledge understood	It is known to which degree the knowledge demands can be satisfied with existing knowledge, and what the gaps are.	**In progress**
- Broad questions formulated	A set of broad questions that reflect the purpose of the research endeavour is formulated.	**No**
- Stakeholders in the loop	The broad questions are agreed with the key stakeholders.	**N/A**
Prerequisites:	*Recommended achieved states of other project aspects.*	
- Research Goals: Demanded	Prerequisite status: 0% progressed. The demand for social data science artefacts that addresses the goals is established.	**Consider**
- Stakeholders: Involved	Prerequisite status: 0% progressed. The stakeholder representatives are contributing towards their responsibilities through engagement in the project.	**Ingore (N/A)**

Figure 4.2: Example alpha and state scorecards taken from the spreadsheet representation of the Social Data Science Scorecard Deck.

the state's progress. The percentage values of the state progress (*30%*) and of the alpha progress (*7%*) are calculated automatically and automatic colour-coding is applied.

As the content of the scorecards were revised throughout the work on this thesis, I stored the scorecards content as a JSON file and developed a Google App Script app that could generated an automated spreadsheet from the data. Storing the contents of the Scorecard Deck as a separate JSON file is valuable in itself, as this allows for development of alternative interfaces.

4.3 Content of the Scorecards

This section examines the individual scorecards that have been developed as part of this research. It is focused on justifying the content of those scorecards with respect to the evidence from the case studies the literature. Most relevant bits of evidence are discussed in detail elsewhere in this thesis and thus only summarised and cross-referenced here. In rare cases when the relevant evidence does not fit the overarching narrative of other chapters, it is provided in this section in full. The discussion of the scorecards that are directly adapted from the original SEMAT Essence model is relegated to the appendices as it does not report on substantive research work, mostly outlines insignificant changes to the original cards and, by its nature, is very repetitive (see Appendix A).

As much as the original SEMAT cards (see Section 2.2), the new scorecards represent the progression of the project core aspects (alphas) through their consecutive states. It is worth noting that in practice such progress is iterative – data acquisition and selection criteria evolve over time and involve trial-and-error, findings motivate new research questions, new data sources get added and the infrastructure has to be adjusted accordingly. Yet, one still can elicit observable states within each iteration, and while an alpha can be returned to one of its previous states before completing the iteration (e.g. if quality assurance of data shows that the acquired data are not of sufficient quality for a meaningful analysis, data would have to be re-operationalised and re-acquired), the alpha would still have to go through the previously achieved states again.

The following discussion tackles the developed scorecards by going through the alphas one-by-one. As the alphas represent *concurrent* project aspects and thus are tackled in an arbitrary order. Within each alpha, the individual state scorecards are discussed. The discussion thus starts with the first state of the "Data" alpha.

4.3.1 Data

4.3.1.1 State 1. Envisioned

The idea that data must first be envisioned is inspired by De Vaus's (2001) discussion of research design. He argues that the issue of which subjects and which of their features data must represent precedes that of particular data acquisition practices and data analysis methods. From this perspective, research design is a combination of the broad research questions (given as a prerequisite in Table 4.1) and the initial vision of data that would support answering those questions conclusively. Following De Vaus's logic, the initial vision of data is vital regardless of whether data are generated in a controlled environment or are naturally occurring, whether they are qualitative or quantitative, big or small. Therefore, envisioning data is as vital for social data

science projects as to more traditional forms of research.

The conditions of this state are derived from a combination of the fieldwork evidence and De Vaus' (*ibid.*) thinking on research design. Under such views, social research is often interested in studying particular aspects (*primary attributes*) of a particular real-world entities (*primary research subjects*). For example, in the InfoMigrants and ShakespeareLives evaluation case studies, among other research tasks, studied the strength of online audience engagement (the attribute) shown by the InfoMigrants and ShakespeareLives audiences (the subjects). In line with De Vaus' logic, the strength of the conclusions that we could make about their levels of engagement (whether they were low, high or normal) depended on our ability to identify other comparable programmes and initiatives (*secondary research subjects*), and to make comparisons with the levels audience engagement those managed to provoke.

In the Shakespeare Lives evaluation, we faced a principle inability to do so, as the that cultural programme was unique in many regards. This caused significant struggles for interpretation of findings, as discussed in Section 3.2.5.3. In the subsequent InfoMigrants evaluation, we did identify comparable initiatives. However, since no one of the compared initiatives performed exactly the same activities as InfoMigrants, we had to elicit their *secondary attributes* (qualitative differences such as focus on specific group of immigrants, provision of specific types of services, etc.) that could interfere with the primary attribute of our interest (see Section 3.4.1.2).

DATA: ENVISIONED	*There is a tentative idea of what kind of data are required for the project needs.*
Primary subjects of research interest determined	The real-world entities that are the subject of the research questions and that must be represented in data are determined.
Secondary subjects of research interest determined	If relevant, secondary groups of real-world entities that should be covered by data for the needs of baseline / benchmark comparison are determined.
Primary attributes of research interest determined	The characteristics of the objects of interest, of their relationships and of their interactions that are of direct concern for the research questions and that must be represented in data are determined.
Secondary attributes of research interest determined	If relevant, secondary characteristics of real-world entities that should be studied in conjunction with the primary ones are determined.
Research Questions: Outlined	*The research questions are formulated as a broad framework that sets the direction for the project.*

Table 4.1: Data: Envisioned. Conditions and prerequisites.

4.3.1.2 State 2 Operationalised

The fieldwork shows how close *examination of data sources* and a corresponding *refinement of research questions* (reflected in both prerequisites and conditions in Table 4.2) help to progress from a vision of data to an operational understanding of the required data. In the Hit List production example, the particularities of each online platform and of their data acquisition modes helped to establish the features that must be represented in the data and the level of granularity of the acquired data records (i.e. *data demands*), as well as the precise criteria of data acquisition formulated as queries to a platform's API or feed (i.e. *data boundaries*) (see Section 3.3.3.1).

Table 4.2 also highlights the role of the compliance strategy in operationalisation of data demands. It is worth noting that in the studied projects compliance never actually limited how a team had to operationalise data. However, a *match between data demands and compliance* had to be ensured when studying the Dark Net. It resulted in an elaborate self-evaluation of ethical and risk management compliance of the team's data collection strategy with a complex risk-assessment framework suggested by Martin and Christin (2016).

DATA: OPERATIONALISED	*Precise specification of which data should be acquired and used in the project is in place.*
Practicalities of data sources considered	The evaluation of potential data sources informs the details of what the practically available data may consist of.
Research questions considered	Any further specifications to the research questions – especially in regards to the relevant features that the data must contain – are considered.
Data demands outlined	Decisions are made on how much data are required, of what types and for which forms of processing.
Boundaries established	The time-, spatial and other boundaries on the data to be acquired are precisely formulated.
Compliance strategy factored in	The compatibility of using the outlined data with the compliance strategy is confirmed.
Data Sources: Evaluated	*The merits and limitations of each source and of their combination are assessed to formulate a firm selection of data sources.*
Compliance: Strategised	*The team members share a clear understanding of how they can achieve ethical and legal compliance.*
Research Questions: Refined	*The questions are redefined and specified to an operational level.*

Table 4.2: Data: Operationalised. Conditions and prerequisites.

4.3.1.3 State 3. Acquired

After the data demands and boundaries are defined to an operational level, data can get acquired from *active data sources* into *a working project infrastructure* (hence the prerequisites in Table 4.3). The specific conditions suggested for this alpha follow the discussion in our chapter on data curation (Voss et al., 2016). The fieldwork positively confirms the importance of these conditions. For example, new restrictions to the Facebook Graph API that were faced in the second year of the InfoMigrants evaluation project show that data cannot always be easily reacquired as it may seem (see Section 3.4.1.1). The Shakespeare Lives evaluation project shows that keeping an extensive "paper trail" of data characteristics may be the best way to track provenance of the research decisions and thus ultimately have confidence in the findings (see Sections 3.2.6.2 and 3.2.5.3). An appropriate data storage structure was especially important for the Hit List production as some of the data pre-processing had to be applied automatically to newly arriving data for the production cycle to finish on time (see Section 3.3.6.2).

The importance of storing the *original* data rather than only a cleaned and processed version was reaffirmed in one of the subsequent studies of cryptomarkets. As data were acquired through scraping HTML pages to JSON files, the acquisition outcomes relied on our knowledge of the pages' DOM trees and on the consistency of those DOM trees over time. Sometimes we would randomly encounter special cases when the DOM structure was different – e.g. banned cryptomarket traders had an additional HTML element on their profile page, which we did not know about for the first several months of scraping and thus had not factored it in the scraping procedure. The only way for us to repair this omission in the earlier scraping iterations was to apply the reviewed scraping approach to the original raw HTML source files which we luckily stored alongside the processed JSON.

4.3.1.4 State 4. Quality Assured

The name of this state could be interchanged with such synonyms as "Cleaned" and "Pre-processed", however "Quality Assured" sound more inclusive than the former and shows a clearer intention than the latter (at the end of the day, the line between "pre-processing" and "processing" is quite thin). The contents of the state's scorecard (see Table 4.4) mostly follow the experience of Hit List collating. Before the analyst could perform the core data analysis, *selection* had to be applied to Twitter data and *aggregation* had to be applied to data from all studied platforms bar Facebook (see Section 3.3.3.1). As these quality assurance activities were done to facilitate the core analysis task, they were designed with this task in mind (hence the *"Analysis Methods: Selected"* prerequisite). On the other hand, the lack of bringing the Twitter data to a *unified schema* with data from other platforms was what caused low visibility of Twitter

DATA: ACQUIRED	*The data are acquired, fully documented and appropriately stored.*
Original data kept safely	If possible, the originally acquired data are fully preserved.
Data storage mode chosen	Data storage mode (flat files, database, etc.) is chosen.
Data structure in place	Appropriate data structure, convenient for the team, is developed (e.g. folder structure for flat files, schema for relational database).
Security concerns addressed	The sensitive data are sufficiently protected to the levels that satisfy the compliance requirements.
Data fully documented	The data are supported with human-readable documentation and machine-processable metadata that describe the data and track their provenance.
Back-ups available	The data are stored redundantly with an agreed redundancy factor.
Compliance: Secured	*The compliance resources required for the project are secured to the degree that the team can fully engage in the main body of the research work.*
Data Sources: Active	*A data acquisition procedure is in place and successfully acquires the required data.*
Infrastructure: Operational	*The infrastructure is in use in an operational environment.*

Table 4.3: Data: Acquired. Conditions and prerequisites.

data for the production team compared to the other studied platforms (see Section 3.3.5.2).

The importance of *dealing with data veracity issues* seems self-evident. The Shakespeare Lives project showed that, interestingly, even interim project data created by the research team poses veracity issues. I encountered those when visualising human data coding (see Section 3.2.6.4). Even though the Excel spreadsheets that the researchers had used to annotate the acquired social media postings had data validation rules applied, somehow several researchers still managed to make spelling mistakes in the names of the variable levels. Furthermore, it was quite common among the researchers to accidentally skip a question and thus create a missing value. Finally, one researcher once changed the ordering of the columns in her spreadsheet which caused my code, hardwired to a precise data schema, to "swap" values of two variables in the resulting visualisations. This bug was easy to repair, but hard to *notice*. Luckily, I corrected the mistake before the researcher had time to use wrong plots for her analysis.

The need to *take compliance measures* when performing quality assurance of data was never high in the studied projects. I decided to include the respective condition as a logical next step from

the compliance-related conditions in the previous states and as an extra precautionary measure.

DATA: QUALITY ASSURED	*The data are brought to the required level of veracity, completeness and usability.*
Data veracity assured	Missing values, unreliable data entries and data glitches resulted from the imperfections of the data acquisition procedure are dealt with.
Selection criteria applied	Criteria for what data to consider for further analysis are in place and applied to the data.
Data aggregation performed	If relevant, individual data records are aggregated.
Data record structuring performed	If relevant, unstructured data records and brought to a defined schema.
Compliance measures taken	The data records and elements that violate the compliance requirements are dealt with (e.g. removed or appropriately anonymised).
Analysis Methods: Selected	*The final list of research methods is compiled.*

Table 4.4: Data: Quality Assured. Conditions and prerequisites.

4.3.1.5 State 5. Utilised

In its final state, data are utilised for the project needs. The conditions suggested for this state (see Table 4.5) are trivial. They add an extra layer of confidence in that the measures that should have been already taken to bring the "Data" alpha through some of its earlier states (such as data documentation and preservation) are in place. As such, they are informed by the same episodes in the fieldwork. They also suggest *archiving* the data where possible to strengthen research reproducibility and to support future studies (Potthast et al., 2016).

DATA: UTILISED	*Data are put to use for the goals of the project.*
Full value extracted	The potential of the data to inform the research questions is reached.
Interpretability ensured	Descriptions of the data and of their provenance are used to interpret the findings.
Data archived	If possible, the data are archived for further utilisation.
Analysis Methods: Executed	*The methods are utilised for the needs of the project.*

Table 4.5: Data: Utilised. Conditions and prerequisites.

4.3.2 Data Sources

4.3.2.1 State 1. Identified

As the Essence of Social Data Science (see Section 4.1) treats data sources as a resource for research, the Social Science Scorecard Deck suggests that envisioning data is a prerequisite to the "Identified" state of the "Data Sources" alpha (see Table 4.6). Putting "Data: Envisioned" as a prerequisite does not lead to assuming a particular linear process of doing social data science where the search for appropriate data sources strictly follows designing the research. In fact, the example of the Hit List production suggests that social data science projects can be first motivated by an opportunity to harness data from some source rather than by any preconceived research question. Arguably, the same may happen in academic work as well.

Rather, this prerequisite tries to convey the idea that it only makes sense to talk about the appropriateness and completeness of the identified list of candidate sources when there is a defined purpose for them. This corresponds well with the aforementioned idea of Gitelman (2013) that data are not data unless they are imagined as such. To follow on the Hit List production example, until the initial idea of using online platforms to learn "what's buzzing" turned into a vision for data on **popularity** of stories among the **UK** public **online** (see Section 3.3.1), it was not possible to claim with confidence that, say, Google Trends was a relevant data source while VK.com was not. By contrast, in the Shakespeare Lives evaluation project, VK.com was very relevant for *demographic* reasons (see Section 3.2.4.1), while Google Trends was not since the analysis focused on audience engagement on social media and therefore search counts were not a *relevant data type* (see Section 3.2.1.1).

Be it to cross-validate findings across audiences in different countries (Shakespeare Lives) or to cover a broad spectrum of popular topics from hard news to viral social media content (Hit List), both these projects had to identify and employ a multitude of sources to allow for *triangulation and complementation*. Data sources with *privileged access* played a key role in the two studied evaluation projects. For example, when evaluating InfoMigrants, some members of our team were given editor rights to their public Facebook pages so that we could access Facebook Analytics.

The data sources and the resulting data, according to the definition of the alpha, are primarily (born) digital but not necessarily so. As the fieldwork shows, social data science projects sometimes use *traditional social science data sources*: interviews, surveys, focus groups, etc. For example, these were heavily incorporated into the work on the "Cultural Value" Strand of the Shakespeare Lives study (see Section 3.2.1.2). Using traditional social science data sources may involve trade-offs in effort and expenses with the new data sources. Also, thinking

about traditional and data-driven components of the research simultaneously (be it data sources, analytic methods, analysis infrastructure, etc.) can lead to a greater synergy between the two. For example, the traditional methods may prove to be especially valuable for finding explanations of the social phenomena that can be observed using the naturally occurring digital data.

DATA SOURCES: IDENTIFIED	*The possible relevant data sources are identified and a preliminary list of appropriate data sources is compiled.*
Relevant types of data considered	The relevant types of data are imagined and the sources for them are identified. E.g., for user-generated online data, data types could be short posts, long posts, threaded discussions, pictures, videos, etc.
Relevant demographics considered	The sources that are populated with data by / on relevant geographical-, age- or other demographic groups are considered.
Available data sources with privileged control / access considered	Opportunities for privileged access to relevant data that are not in the public domain are considered. Examples may include website visit statistics, customer relationship data, etc.
Traditional sources considered	Traditional social science data acquisition sources such as interviews and focus groups are considered as a data source.
Complementation and triangulation assured	The identified data sources allows acquiring data that cover different aspects of the studied problem and cross-validate each other.
Data: Envisioned	*There is a tentative idea of what kind of data are required for the project needs.*

Table 4.6: Data Sources: Identified. Conditions and prerequisites.

4.3.2.2 State 2. Evaluated

The Social Data Science Scorecard Deck puts a huge emphasis on evaluation of potential data sources as their diversity means that they have varied limitations and particularities. As suggested by the fieldwork, these limitations and particularities can, in turn, have profound consequences for methodology and for the employed infrastructure. This leads to quite an elaborated system of conditions in the respective scorecard (see Table 4.7).

The issues of *degree of access* and *level of control* are related to each other. It is tempting to think about them in terms of a dichotomy between the internal (and thus fully-accessible and well-controlled) versus external, problematic data sources. The fieldwork, however, suggests that this is not the case and, in fact, different modes of acquisition allow for a trade-off between the two. While the commercial tools employed in the Shakespeare Lives evaluation study to acquire data from Twitter and Instagram provided privileged access to the data, the control over

the export functions of those systems and, therefore, over the ability to subject the accessed data to our team's own analysis methods, was quite limited (see Section 3.2.6.1).

Different data sources may lead to different levels of *data veracity* that have to be considered. As already mentioned, two of the studied projects (evaluations of Shakespeare Lives and InfoMigrants) used both traditional sources of data alongside digital platforms. The veracity issues obviously differ between these two groups. However, even within digital platforms, data veracity strongly depended on the available modes of acquisition. For example, as there were restrictions to accessing the Facebook Graph API in the second year of the InfoMigrants evaluation (see Section 3.4.1.4), some limited scope scraping had to be performed. Since Facebook's website is highly dynamic, the scraping outcomes were not always as expected and had to be partially manually repaired, which would have been problematic had the scope of scraping been higher.

Considering the *mode of acquisition* thus becomes an overarching issue for all the ones discussed above. Another reason why an acquisition mode matters is that it may pose its own *compliance* restrictions that must be considered. For example, some platforms strictly prohibit scraping. What particular ramifications not adhering to this rule may have and whether it changes depending on the scope of scrape is a separate question – yet it is a question a researcher should be prepared to answer.

The idea of source-wise biases is informed by Ruths and Pfeffer (2014) who outline how different platforms attract varying demographic groups and, through the platform mechanics, foster different kinds of behaviour; although the issue of platform-specific bias is tackled in numerous literature (Sloan, 2017; Cihon and Yasseri, 2016; Morstatter et al., 2014). The fieldwork strongly supports this idea, with the systematically repeating difference of dominating topics across the studied platforms discovered during Hit List production being a primary example 3.3.3.2. Behavioural biases caused by the platforms were also one of the reasons why, within the Shakespeare Lives evaluation, data from Twitter and Sina Weibo platforms were acquired with keyword-based criteria and were studied through human data annotation as individual postings, while data acquisition for other platforms was focused on particular accounts and the data were studied through online ethnography (see Section 3.2.4.2).

Having feasible *infrastructural requirements* to data acquisition from each data source were critical in every studied project. For example, the work on the Hit List was only possible since there were multiple servers located in different universities and thus connected to different university networks in the team's possession. The network-independence and physical dispersion of the servers were crucial to achieve a sufficient level of redundancy in acquiring data from

the sources that required continuous access, such as the Twitter Streaming API and the Google Trends feed. It is worth noting that the relationship between the data sources and the infrastructure can go both ways – i.e. architecture may be selected to meet the data acquisition requirements. As will be shown further, this is represented in the "Infrastructure" alpha (see Appendix A.3.1.1).

Finally, the *mapping to the research questions* condition is left in mostly to guarantee that this crucial match, which has already been discussed above (see Section 4.3.2.1) is maintained.

DATA SOURCES: EVALUATED	*The merits and limitations of each source and of their combination are assessed to formulate a firm selection of data sources.*
Level of control estimated	The ability to reliably acquire the expected data from each data source is assessed.
Degree of access considered	It is identified which part of the potentially relevant data are actually available for acquisition from each source.
Data veracity estimated	The factors that can undermine veracity of the acquired data are estimated for each source.
Sources of potential bias identified	Potential sources of bias / unrepresentativeness of data coming from each source are identified.
Modes of acquisition considered	For each source, the available modes of data acquisition are assessed for their costs, reliability and ability to return specifically the relevant data.
Infrastructural requirements considered	The feasibility of infrastructure for data acquisition, storage and analysis are assessed for each source.
Preliminary mapping to the research questions established	There is a preliminary understanding of how each of the broadly outlined research questions can benefit from data by each source.
Compliance: Considered	*It is identified which ethical and legal concerns the project raises.*
Infrastructure: Architecture Selected	*Architecture has been selected. It addresses the key technical risks and any applicable organisational constraints.*
Research Questions: Outlined	*The research questions are formulated as a broad framework that sets the direction for the project.*

Table 4.7: Data Sources: Evaluated. Conditions and prerequisites.

4.3.2.3 State 3. Selected

This state represents the *finalisation* of the choices that emerge from identification of potential data sources and evaluation of their merits and limitations (see Table 4.8). For example, as discussed above, the *specialisation* decisions can naturally follow the differences in the represented demographics and in the platform-provoked user behaviour. The *prioritisation* of

sources can be clearly observed in the Shakespeare Lives project. The Chinese researcher used data from two platforms – Seana Weibo and Twitter – for content analysis. Treating Twitter as a secondary source in this scenario allowed her to be a bit less meticulous in iterating through data acquisition criteria, which was very liberating from the logistics perspective (see Section 3.2.7.1 details). On the other hand, not neglecting Twitter completely allowed to satisfy the *operationalised data demands* in terms of quantities of data for analysis.

The scorecards also reminds that some of the identified and evaluated data sources may be completely *discarded*, yet it suggests that there must be "convincing reasons" for that. The experience of the case studies shows that if a source has been identified as clearly bearing relevant data, a research team would do their best to employ this source in some shape or form, even if a convenient mode of access to data was absent or the available data were incomparable with data from other sources and not perfectly tailored to the research questions. As has been shown above, instead of simply rejecting a source, the research teams in the studied projects would more likely assign a lower priority to such a source and consider its particularities when interpreting findings. This seemingly "stubborn desire" to keep the previously selected data sources in the scope of a project actually is not by itself methodologically harmful and tends to make the projects stronger. Even if some data sources are especially problematic, it does not that others are at all not – at the end of the day, most social data science sources bear naturally occurring data, which are inherently not tailored specifically to the research questions. Having the need to deal with different sets of limitations and methodological pitfalls for different sources leads to better triangulation of findings and to levelling out the systematic noise in the data related to the properties and imperfections of each individual data source.

DATA SOURCES: SELECTED	*A firm decision on the use of each data source is made.*
Specialisation decisions made	It is decided if each data source will be used to acquire data on a particular aspect of the studied problem.
Prioritisation decisions made	It is decided whether particular data sources will be used as primary and others as secondary.
Discard decisions made	If there are convincing reasons, it is decided if some of the potential data sources will not be used at all.
Selection finalised	There is a finalised selection of data sources to use.
Data: Operationalised	*Precise specification of which data should be acquired and used in the project is in place.*

Table 4.8: Data Sources: Selected. Conditions and prerequisites.

4.3.2.4 State 4. Supported

The "Data Sources" alpha progresses to this state when everything is ready for data acquisition (see Table 4.9). Having an *obtained access* to the sources is thus a self-evident condition. The conditions of *understanding risks* and *having alternatives* are related to each other, as the alternatives have to be put into correspondence with the potential problems that can occur. The aforementioned infrastructure architecture that allowed to acquire the Twitter 1% sample redundantly is a good example. If the main project server had crashed or there has been a power/network outage, the data would still have been acquired in a different university. A risk that a data acquisition mode associated with any particular platform would at all become unavailable was diversified by relying on data from multiple platforms.

It is of course true that some risks cannot be foreseen – and even if they can, they seem sufficiently unlikely and a temporary interruption in data acquisition would not pose a huge challenge anyway. In such cases, it may not be worth to design alternative data collection procedures in advance. For example, this was the case with the terminated access to the Facebook Graph API in the second year of the InfoMigrants evaluation project – this had not seemed as a likely scenario at least until the Cambridge Analytica scandal emerged. It is up for a research team who uses the Scorecard Deck to figure out which situation is more applicable to them.

Regarding the prerequisites, having an *infrastructure ready* to acquire the data (or at least to store it if the data are purchased) and *securing the compliance* to access the data sources (i.e. getting relevant ethical permissions, licenses, data access rights) seem to be self-evident. Having a *collaborating team* may seem to an overkill – in principle, data acquisition may be started by a sole data scientist who plans to form a team to analyse these data later. Yet, the fieldwork suggests that having collaborating team members at the data acquisition stage is hugely beneficial, as the data selection decisions are likely to be made iteratively and, while sometimes they can be made post-factum (e.g. while working on Hit List – see Section 3.3.4), it is not always possible. Yet again, the collaborative process of formulating Twitter data acquisition criteria in the Shakespeare Lives project is a good example of that (see Sections 3.2.6.1 and 3.2.7.1).

4.3.2.5 State 5. Active

Many prerequisites and conditions of the "Data Sources: Active" state are logical next steps of those discussed alongside the earlier states of this alpha (see Table 4.10). This includes the issues of *compliance, infrastructure, maintaining redundancy and monitoring*. The last condition – *a warrant of continuous access* – is left specifically for cases of using modes of acquisition that may have an upper bound on the time of their use due to the terms of conditions. In the studied projects, such modes were presented by the aforementioned commercial social media analysis

DATA SOURCES: SUPPORTED	*The conditions required to actually start data acquisition from each source are met.*
Access obtained	For each data source, there is a reliable and feasible way to access the data (API, scraping, data vendor, etc.).
Risks understood	Risks to data acquisition process associated with the chosen way to access data are understood.
Alternatives being available	If possible, an alternative data acquisition process is designed that can be invoked if the main plan fails.
Compliance: Secured	*The compliance resources required for the project are secured to the degree that the team can fully engage in the main body of the research work.*
Infrastructure: Ready	*The infrastructure has been accepted for deployment in a live environment.*
Team: Collaborating	*The members of the team effectively collaborate with each other.*

Table 4.9: Data Sources: Supported. Conditions and prerequisites.

tools employed in the Shakespeare Lives project.

DATA SOURCES: ACTIVE	*A data acquisition procedure is in place and successfully acquires the required data.*
Data being acquired	The required data become available to the project.
Monitoring in place	Continuous monitoring that right data are acquired when expected from each source is in place.
Redundancy achieved	The infrastructure allows for redundant acquisition in case of partial infrastructure failure.
Continuous access warranted	Accessibility of the required data throughout the whole period of acquisition is assured.
External changes monitored	The changes to license agreements, APIs, acquisition tools and other parameters out the project control are monitored and accounted for.
Infrastructure: Operational	*The infrastructure is in use in an operational environment.*
Compliance: Maintained	*The team works on the project while reactively and proactively maintaining compliance.*

Table 4.10: Data Sources: Active. Conditions and prerequisites.

4.3.2.6 State 6. Utilised

The last state of the "Data Sources" alpha, which is achieved after *completion of data acquisition*, does not, in principle, require any additional conditions. Yet, a condition of *assuring possibility*

of re-use is added on top. This mostly relates to the infrastructural component of data acquisition – for example, the Twitter data acquisition for the Hit List was performed with the tools and on servers that had already been employed to connect to the Twitter Streaming API by some of the data scientists on the team. However, this also relates to the specific practices and procedures of data acquisition. For example, in the Shakespeare Lives project data acquisition had to be re-invoked for each round of three rounds of reporting. When, after the first round, a new project manager joined the team, she could reuse the data documentation standard developed by the first project manager, which significantly eased keeping track of the data acquisition progress (see Section 3.2.6.2).

DATA SOURCES: UTILISED	*Data from each source have been successfully acquired.*
Data acquisition complete	The data acquisition process is complete.
Possibility of re-use assured	If possible, there is an established way to re-invoke the data acquisition from each source – preferably at a lower cost and with a shorter notice.

Table 4.11: Data Sources: Utilised. Conditions. No prerequisites specified.

4.3.3 Research Questions

4.3.3.1 State 1. Outlined

The Scorecard Deck distinguishes research questions posed in a broad form of more open questions formulated in terms of the subject domain and their subsequent refined version posed in terms compatible with the imagined data and methods. For the first state of the "Research Questions" alpha to be achieved, the former is sufficient (see Table 4.12).

Translating the *outlined research goals* into specific *knowledge demands* was a crucial bit for all the studied research projects, yet the difficulty of it was different from project to project. For the Hit List production, this translation was quite straightforward, as the goal of creating a weekly chart of stories based on their popularity online in the UK contains a question in itself. By contrast, for the Shakespeare Lives evaluation, this translation was significantly more difficult. The goal of assessing the online performance of a cultural programme does not specify neither the particular subjects of interest neither the success criteria. To make this translation, the project investigators had to rely on their past experience of similar evaluations, on their understanding of the priorities of a soft power organisation and on further *communication with the stakeholder* – i.e. the British Council. As can be seen, the *list of broad research questions* for Shakespeare

Lives evaluation (see Section 3.2.4) is, while not straight-away operational, really manages to capture what should be known by the end of the project.

The need of doing the background research and *understanding the scope of existing knowledge* was not something that was really encountered in the key case studies undertaken as part of the fieldwork, as the studied projects aimed to generate knowledge for very narrow and specific purposes. Rather, this condition naturally emerged from the general understanding of research process. Subsequently, when evaluating the Scorecard Deck while studying the crypto-criminal marketplaces on the Dark Net, this condition proved to be very relevant, as the background knowledge brought to the team by the project's principal investigator was crucial to correctly pose the research questions on the resilience of such marketplaces to law-enforcement intervention in terms of the social structures and relationships between buyers and sellers that these marketplaces facilitate.

RESEARCH QUESTIONS: OUTLINED	*The research questions are formulated as a broad framework that sets the direction for the project.*
Demand understood	The research needs and goals can be transalted into specific knowledge demands.
Scope of existing knowledge understood	It is known to which degree the knowledge demands can be satisfied with existing knowledge, and what the gaps are.
Broad questions formulated	A set of broad questions that reflect the purpose of the research endeavour is formulated.
Stakeholders in the loop	The broad questions are agreed with the key stakeholders.
Research Goals: Demanded	*The demand for social data science artefacts that addresses the goals is established.*
Stakeholders: Involved	*The stakeholder representatives are actively involved in the work and fulfilling their responsibilities.*

Table 4.12: Research Questions: Outlined. Conditions and prerequisites.

4.3.3.2 State 2. Refined

The fieldwork suggests several conditions the research questions have to met to be considered refined. First, the questions have to be *scoped*. For example, in the Shakespeare Lives case study, the broad research question on the sentiment of response to the cultural programme, the notion of response had to be scoped. Some of the considered issues were whether the response had to come in a form of separate social media postings or as a reaction (e.g. liking) on other posts; whether the response should be about the cultural programme as a whole or about its particular events; in the latter case, whether there must be evidence that the commenting person understands that the event is part of a wider Shakespeare Lives campaign. This scope was actually refined iteratively

within the project (see Section 3.2.4.3), yet understanding it and its implications on each stage of research was crucial to framing the emerging findings correctly.

Specifying variables and their values for the coding frameworks (see Section 3.2.5) employed in the Shakespeare Lives and InfoMigrants evaluations is a great example of *extracting specific questions*. The frameworks are in correspondence with the nature of the *envisioned acquirable data* (isolated social media postings of topic of Shakespeare Lives / isolated articles on the InfoMigrants website) and are to be used in the suggested mode of analysis (content analysis). The specific research questions can thus be asked in terms of the distributions of the coded variables and of their relationships (see Section 3.2.6.4). The ability of the team to answer those questions has to rely heavily of *common understanding* shared by the team members. For example, in the Shakespeare Lives project, one of the key questions the researchers had to answer in regard to every analysed post was whether it acknowledged the "values of Britain" that the British Council aimed to promote. Understanding what such an acknowledgement must constitute of was an involved process (see Section 3.2.5.1).

The coding variables discussed above quite straightforwardly informed the *proxies and metrics* for answering the refined research questions. Since each variable had a limited number of possible values, demonstrating the empirical distributions of those values was sufficient, provided that a researcher gave an interpretive commentary alongside representation of those distributions (see Sections 3.2.6.4 and 3.2.5.3). In other cases, when the involved analysis included reducing the data to a particular metric, its choice could be debated. For example, the metric used by the analyst to measure a topic's popularity on Facebook when producing the charts for Hit List show – a weighted sum of comments, likes and shares – did not seem like the best choice to at least one producer on board who would have preferred a plain, unweighted sum instead (see Section 3.3.5.1). Yet, what is important in the context of this particular state of the "Research Questions" alpha is that the principle possibility to derive a metric in response to the posed question from the data was there.

4.3.3.3 State 3. Matched

This state of the "Research questions" alpha in the Scorecard Deck plays a bridging role and serves as a reminder for a research team to reflect on their research decisions thus far and to make sure they are consistent and aligned with each other. The importance of some of the conditions – e.g. of checking the match between the research questions on the one hand and *data* and *compliance* on the other hand (see Table 4.14) – is thus trivial and follows from the discussion of previous states of the alpha.

The notion that refined research questions may lose connection to the original *broad questions* is

RESEARCH QUESTIONS: REFINED	*The questions are redefined and specified to an operational level.*
Scope established	The questions are scoped to be feasible to answer within the project boundaries.
Specific questions extracted	Specific questions answerable with acquirable data and appropriate executable methods are elicited.
Proxies and metrics envisioned	Plausible quantitative and qualitative metrics / indicators useful for answering the questions are designed.
Common understanding ensured	The questions are agreed by the team and, if relevant, the stakeholders.
Data: Envisioned	*There is a tentative idea of what kind of data are required for the project needs.*

Table 4.13: Research Questions: Refined. Conditions and prerequisites.

motivated by some of my prior work, e.g. one reported in Willis et al. (2015). In that project, our team performed social networking analysis for the BBC World Service. While that project did come to a successful end, part of the social network analysis performed by out team was rejected by the stakeholders as not relevant to their needs and to the broad questions that the proposal had promised to answer. Working as a more-or-less independent team of social network analysts and occasionally allowing ourselves to pursue what *we personally* thought were the most interesting aspects of the data was the primary reason for that.

Finally, the issue of relationship between research questions and *research artefacts* is probably best demonstrated by looking at the attempts to use an unsupervised learning algorithm to cluster Google Trends search terms into topics. Indeed, that classifier essentially was designed to answer a particular question of how we could group the search terms by the vocabulary used in related news headlines. While this question corresponded well with the broad question of when two separate themes can be considered one, it was still not the right research question specifically for *producing the Hit List*, since radio show production is inherently *not* agnostic of past experience (i.e. the work on previous episodes – and the contents of those episodes), while such a classifier is (see Section 3.3.4.3).

4.3.3.4 State 4. Answered

The key issue that the Scorecard Deck tries to emphasise in regard to answering research questions is *considering limitations* of the employed methods and data sources (see Table 4.15). While this issue is by no means new or unique to social data science, the discussion of the "Data Sources" and the "Analysis Methods" alphas (see Sections 4.3.4 and 4.3.2) shows just how much their role matters – and how it may be different to that of traditional social science empirical

RESEARCH QUESTIONS: MATCHED	*The refined research questions are re-validated to match the research circumstances.*
Data match ensured	It is ensured that the metrics / indicators required to answer the research questions are actually derivable from the data.
Compliance match ensured	The research questions are validated to be compliant with the formulated compliance strategy.
Broad questions match ensured	It is ensured that the specified questions still correspond to the original broad questions.
Artefacts match ensured	It is ensured that answering the research questions can help to progress towards the envisioned research artefacts.
Data: Operationalised	*Precise specification of which data should be acquired and used in the project is in place.*
Compliance: Strategised	*The team members share a clear understanding of how they can achieve ethical and legal compliance.*

Table 4.14: Research Questions: Matched. Conditions and prerequisites.

methods of data collection and analysis. For example, in the Shakespeare Lives evaluation, our team encountered tweets that had been almost certainly posted by bots (see Section 3.2.5.2). While such clear cases of spam equipped out team with greater consciousness towards this issue, they also meant that there might be other cases when we would make a false judgement one way or another. The risk of spam influencing the analysis findings is something hard to imagine in traditional social science.

RESEARCH QUESTIONS: ANSWERED	*Through applying the analytical methods to the acquired- and quality assured data, actionable answers to the research questions are derived.*
Findings collated	The findings acquired by different methods and data sources are put in context of each other.
Limitations considered	The limitations and specificities of the used research methods and data sources are used to interpret the findings.
Answers formulated	The answers are provided to the research questions. The limitations to each answer are clearly identified.
Data: Quality Assured	*The data are brought to the required level of veracity, completeness and usability.*
Analysis Methods: Executed	*The methods are utilised for the needs of the project.*

Table 4.15: Research Questions: Answered. Conditions and prerequisites.

4.3.3.5 State 5. Utilised

The last state of the alpha is distinguished from the previous one to remind that simply arriving to answers to the research questions does not end their journey. The answers have to be *(*embedded) into project artefacts – be those numerous project reports and an interactive website for the visualisations as in the case of Shakespeare Lives or as a weekly topic chart and then a radio show based on it as in case of Hit List. Second, the experience of the fieldwork shows that if there was a stakeholder who evaluates the project outcomes (and there arguably almost always is – e.g. in purely academic research that would be peer reviewers), the answers to the research questions would go through iterative refinement even after the team is happy with them. This was especially clearly seen in the Open University projects, since both the British Council in one of them and the InfoMigrants team plus the European Commission in the other, challenged the presented reports and suggested new angles to look at. Finally, each round of answering research questions often led to *extending them* for further research. For example, the online ethnography of the British Council's public pages Facebook and VK.com led to a futher question of how the discussions around Shakespeare Lives evolve on the pages of the British Council's local partners in each country. This question was investigated on the last round of reporting.

RESEARCH QUESTIONS: UTILISED	*Answers to the research questions are in use.*
Full value extracted	The full value of research answers for progressing towards the research artefacts is utilised.
Answers reviewed	Research answers have been externally assessed through stakeholder review of artefact iteration(s).
Questions extended	New round of research questions is motivated.

Table 4.16: Research Questions: Utilised. Conditions and prerequisites.

4.3.4 Analysis Methods

The "Analysis Methods" alpha is the one that is the hardest to conceptualise and represent in the Scorecard Deck, as the underlying concept is an involved one. It is worth reiterating that the alpha is referring to *analysis* methods, as the word "method" by itself is used in a variety of contexts, for example as part of "scientific method"; this understanding would include analysis *per se* as only one of the method's stages, which comes in a lot later than, for example, proposing verifiable (Cohen and Nagel, 2013, p.215) – or falsifiable (Popper, 1959, pp.57–120) – hypotheses. Using the expression "research methods" to refer to different ways of generating knowledge about the world – including those that do not prescribe to a strict scientific process, e.g. grounded theory (Strauss and Corbin, 1990) – aims at a similar scope.

The "Analysis Methods" alpha here refers to the particular techniques that are applied to the acquired data to derive knowledge from them. It is, yet again, in line with the ideas of De Vaus (2001) that research design, acquisition modes and data analysis are agnostic to each other. For example, a researcher may employ experimental research design in that they formulate a hypothesis on an effect of some intervention and then they collect data on two groups of subjects (the treated group and the control group) before and after the treated group was subjected to this intervention. Yet, instead of using statistical analysis methods, which are stereotypically associated with such research, the research may opt for acquiring qualitative data (e.g. through interviews with the participants) and analyse it with qualitative methods (e.g. by identifying the themes emerging in those interviews with open coding).

This is a very powerful idea that is, as the fieldwork shows, very relevant to social data science (see further discussion on the alpha's states for details). Yet, the problem is that a lot of analysis methods are defined with a particular "grander" research method – and, as Janert (2010, p.221–235) argues, even a particular research designs – in mind. For example, even when doing predictive regression modelling in an iterative, data-driven manner without explicitly formulating a prior hypothesis about the form the final model should take, the processes is likely to involve significance testing and, by this, hypothesis testing under the hood.

For this reason, the line between the "Analysis Methods" alpha and other alphas is not always that clear to a potential confusion for a user of the Scorecard Deck. For example, when the "Research Questions" alpha achieves the "Refined" state (see Section 4.3.3.2), if the study is hypotheses-driven, the questions would be formulated in terms of the hypotheses. Likewise, if the "Data" alpha achieves the "Envisioned" state (see Section 4.3.1.1), this would mean that the research is designed. This confusion was indeed noticed in the case study that evaluated the Deck (see Section 5.2.3) – and because of that, the related scorecards were very significantly reworked. Yet, this confusion is likely to not be fully remediable and can be observed elsewhere – not only in the process of using the Scorecard Deck. A great example of that is the use of terms "classifier" in Statistics circles and "hypothesis" in Computer Science circles for the same concept (Wasserman, 2004, p.15).

Hence, instead of aiming at unachievable, the scorecards for the "Analysis Methods" alpha try to foster thinking about the analysis methods in terms of research questions and research design and to make some of the decisions that would otherwise go explicit and unnoticed implicit.

4.3.4.1 State 1. Selected

The first state of the "Analysis Methods" alpha is achieved when the methods are informed to a point of possibility to make an initial selection (see Table 4.17). Informing the methods

with the analysis design was of absolute importance in all the studied project, but most of all in the evaluation of Shakespeare Lives. As discussed in Section 3.2.5.3, the research design of that project was characterised by a mismatch between the overarching *question* of the British Council's interest ("Is our performance good?") and the practical impossibility to find or construct benchmark *data* against which the programme's performance could be evaluated. The use of content analysis as a method to analyse social media posting about the programme – and the way in which this mode was used – was, partially, a response to this mismatch. By analysing big samples of postings through the lens of a detailed, rigid, closed coding framework that nevertheless focused on *qualitative* aspects of the response, it was easier to draw comparisons between responses to the programme in different countries and at different points in time and thus use those as internal benchmarks without being distracted by the a priori varying volumes in reaction to event of different types. This example also shows that the *qualitative analysis methods* from traditional social science – such as content analysis in this case – can be perfectly applicable to new forms of data if align with the research purposes.

Regarding the trade-offs between *labour-intensive and computationally-intensive analysis*, another good example comes from the same project. As discussed in Sections 3.2.5 and 3.2.1.3, analysis of tone of the social media posts was performed both computationally through sentiment analysis using SentiStrength (Thelwall, 2017) and through manual coding of data. The distributions of sentiment shown by both methods were quite similar. From this perspective, using a computationally-intensive method offered a better "value-for-effort". However, if only the sentiment analysis was used, the team would likely not discover the reason why in many postings the tone *had to be* neutral, or "weak" on both positivity and negativity if using the sentiment analysis terminology (the reason is discussed in Section 3.2.5.3). Besides, other categorisations of data that were coded by analysts alongside tone would not be there, so the depth of insight available from data would be questionable. There is no universal solution for what is better or worse in this trade-off – it will depend on the research purposes, size of data and available resources – yet the *option* of using labour-intensive analysis, even if similarly purposed and sufficiently accurate computational analysis methods are available, should still be at least considered. As a side note, the use of SentiStrength was an example of *infrastructure suggesting analysis modes* – if there was no ready software solution for sentiment analysis tailored to social media data, this analysis method would not be tried at all.

4.3.4.2 State 2. Piloted

The fieldwork suggests that refinement of the particularities of the analysis methods in social data science projects happens iteratively in the process of analysis itself, hence the next state of the "Analysis Methods" alpha is achieved when the outcomes of pilot analysis are evaluated to further

ANALYSIS METHODS: SELECTED	*Candidate analysis methods are identified and selected.*
Research design considered	The analysis methods are informed by what kinds of answers the outlined research questions seek and what the envisioned data aim to represent.
Background research done	Experience of similar research projects and relevant methodological advice have been studied.
Relevant techniques considered	The analytical techniques – either specific ones or families of techniques – that correspond with the research questions are identified.
Qualitative analysis methods considered	It is established if pure qualitative methods are appropriate for some of the analysis.
Human/computer trade-offs considered	For the planned analytical pipelines, it is considered which parts would most benefit from being labour intensive vs. computationally intensive and what is the role of human-in-the-loop.
Infrastructural resources considered	Feasibility of setting infrastructure to execute each method is considered.
Triangulation and complementation considered	The analysis methods allow cross-validating and enhancing the findings derived with each other.
Evaluation criteria agreed	It is known how to measure the level of confidence in the results produced by each method.
Research Questions: Outlined	*The research questions are formulated as a broad framework that sets the direction for the project.*
Data: Envisioned	*There is a tentative idea of what kind of data are required for the project needs.*

Table 4.17: Analysis Methods: Selected. Conditions and prerequisites.

inform the analysis methods and the relevant *adjustments are made* (see Table 4.18). An example of refinement that focused on method *validity* can be found in the Shakespeare Lives project, as on the early stages of the project the team had to continuously adjust how the researchers coded social media postings for whether their authors acknowledged the "British values" – i.e. the image of the UK that the British Council aimed to promote (see Section 3.2.5.1). Another example of a validity concern, which was not really tackled, is found in the experience of Hit List production. One of the interviewed producers did not agree with the employed way to chart the posts of news media on Facebook, as, according to her, the *weighted* sum approach taken by analyst to aggregating different engagement metrics was not perfectly valid to the underlying question of which stories caused the largest number of social media interactions (see Section 3.3.5.1).

Ensuring *sufficiency* of the research methods was something that did not require that much

iterative refinement in the observed case studies – presumably due to the past experience of the involved members in similar work. For example, in an earlier Open University research project that I had participated in as a quantitative analyst (Willis et al., 2015) and that was partially funded by the BBC, our sub-team tried to assess the role of BBC's Twitter accounts in the overall Twitter conversations surrounding London Olympics 2012 using the social networking analysis methods. While in our analysis we did manage to arrive at some relevant findings, the question of what could be done differently was left largely unanswered because we did not combine the social network analysis with qualitative content analysis methods that could shed a light on the kinds of Twitter behaviour that caused the most response.

Finally, while in the observed projects I did not encounter situations when the *resource load* chosen analysis modes created was prohibitively high (presumably because the employed methods were relatively computationally cheap to begin with and the labour-intensive methods did not require to be executed within short time frames), there were cases when the involved computational methods required accurate planning and scheduling to be exercised on time – the initial data aggregation and analysis in the Hit List production being the primary example (see Section 3.3.6.2). Should the designed methods have been a bit more computationally involved, it would have most likely required to either simplify them or to invest in higher performance infrastructure.

ANALYSIS METHODS: PILOTED	*The methods are pilot-tested and necessary adjustments are implemented.*
Validity assessed	It is assessed whether the outcomes of method applications are valid for the research question they aim to address.
Sufficiency ensured	It is made sure that all research questions are addressed by at least one method.
Resource load estimated	It is seen if execution of some methods consumes too much computational or labour resources.
Adjustments made	Reflecting on the observed results, the method selection and implementation are adjusted.

Table 4.18: Analysis Methods: Piloted. Condition. No prerequisites specified.

4.3.4.3 State 3. Executed

This state of the "Analysis Methods" alpha is achieved when a certain iteration of analysis of *quality assured data* is finished, with all the required *efforts applied* and the required *outputs stored*. It was normally accompanied by *documenting the the analysis process*. In the Hit List project it was relevant because the Hit List had to be produced on a weekly basis. Therefore, if an analyst got ill or otherwise could not perform their responsibilities on a particular week, without

proper documentation finding a temporary replacement would be hard. In the Shakespeare Lives and InfoMigrants project the documentation was first and foremost used in reporting, as the projects' key stakeholders should have had an opportunity to challenge the team's methodology.

The *evaluation of outcomes* took various forms depending on project and analysis method. In the Hit List production, the goodness of the supervised classifier used to for filtering UK tweets in Twitter data was measured using a procedure typical for machine learning – the classifier was applied to test data not used for training the classifier but containing the country labels. Since the classes were highly imbalanced (i.e. there were much less UK tweets than non-UK tweets), a simple accuracy measure did not make the cut. Thus, precision and recall were used instead. As discussed in Section 3.3.3.3, the derived metrics implied that the analyst could not be confident in the classification outcomes enough to use the filtered list of hashtags without critical judgement. In turn, the quality of this critical judgement – i.e. of additional filtering out the non-UK topics – could not be formally evaluated, so it was instead sanity-checked by the production team.

The *method reproducibility* condition is derived from Goodman et al. (2016). The specificity of the term is to avoid confusion with many other R-dimensions (De Roure et al., 2011), three of which – reproducibility, replicability and repeatability – are known to cause the most confusion (Plesser, 2018). Method reproducibility implies the possibility to rerun the same analysis over the same data to test whether this leads to the same research results – in other words, it requires research workflow preservation (De Roure et al., 2011). The suggested scorecard condition retains the gist of this definition.

Assuring method reproducibility, in principle, implies that methodologically stronger forms of reproducibility can also be tested for, provided that there is sufficient documentation of the analysis process (which is warranted by a different condition). By combining a preserved workflow with an informed judgement on what each of its steps achieves and how it is motivated, it is possible to configure other similarly purposed research pipelines, for example, to test whether the specific implementations of the analysis methods or the specific analysis techniques chosen influence the analysis outcomes. For this reason, no separate conditions concerning other forms of reproducibility are listed in this alpha.

Finally, it is worth noting that the interest in reproducibility beyond that of method reproducibility may often be limited in social data science as many social data science projects do *descriptive* rather than *inferential* research (see De Vaus (2001) for further discussion of the difference) – new forms of data allow to characterise researched subjects in an unprecedented manner and thus are valuable for reach descriptions. Descriptive research was at the core the studied project as well and took form of characterisation of topic popularity in the Hit List project and

characterisation of engagement in the Shakespeare Lives and InfoMigrants evaluation projects. Even when inferences *were* derived in the form of recommendations for stakeholders in the latter two projects, those were programme- or initiative-specific and not meant to be necessarily applicable to larger populations (i.e. to other similar programmes and initiatives should they ever occur).

ANALYSIS METHODS: EXECUTED	*The methods are utilised for the needs of the project.*
Effort and computations applied	All the necessary human efforts and computations that execute each method are performed.
Outputs stored	The intermediary and final outputs of the method application are reliably stored.
Outcomes evaluated	The team members know how confident they are in the outcomes of each method.
Process documented	The process of method execution is fully documented and relevant metadata are generated and preserved.
Method reproducibility assured	It is possible to rerun the performed analysis and test wether same data produce same outcomes.
Infrastructure: Ready	*The infrastructure has been accepted for deployment in a live environment.*
Team: Performing	*The team is efficient and effective at progressing its work.*
Ways of Working: Employed	*The conditions required for the work to start are met.*
Data: Quality Assured	*The data are brought to the required level of veracity, completeness and usability.*

Table 4.19: Analysis Methods: Executed. Conditions and prerequisites.

4.3.4.4 State 4. Extended

As much as research questions, new analysis methods also got *proposed* after *critique* on the last iteration was formulated (see Table 4.20). For example, in the InfoMigrants evaluation after the first year it became clear that the interpretation of the quantitative findings on the engagement with the programme that our team provided did not add much value to what the InfoMigrants own marketing team could obtain with commercial software – and, crucially, it was not very revealing about what had been done good and what not so good. Hence, in the second year the team went for discourse analysis of the comments on the InfoMigrants Facebook pages in addition. Interestingly, a new research question was also posed alongside, since it could be conveniently addressed by discourse analysis instead – the team was tasked to access the *quality of debate* (Habermas, 1991) unfolding on those pages.

ANALYSIS METHODS: EXTENDED	*Next iteration of analysis (within or outwith the project) is methodologically informed.*
Critique formulated	The team understands the overarching methodological issues and limitation of their analytical work.
New methods proposed	New analytical methods that can remedy the limitations are proposed.

Table 4.20: Analysis Methods: Extended. Condition. No prerequisites specified.

4.3.5 Artefacts

4.3.5.1 State 1. Outlined

The first state of the "Artefacts" alpha is achieved when there is understanding of the form that project artefacts can take, but not necessarily what their contents would be. For example, the artefacts outline may take a form of agreement between the project team and *the key stakeholder* (either the funder or the beneficiary). In the Shakespeare Lives evaluation project, such an agreement included the description of responsibilities delegated to the three project strands (see Section 3.2.1) in terms of *type of desired artefacts* – language reports, visual analysis reports, a Cultural Value report, infographics and data visualisations, and a single web-site that would serve as a platform to host the other artefact types. The Wire Free Production company concluded a similar agreement with the BBC Radio 5 live in terms of delivering a weekly episodes of the Hit List show. Each of the artefacts mentioned in those two agreements reflected a particular *research goal* – be that informing British Council on the public engagement that their programme had provoked in different countries or informing the British audience on the most discussed stories.

4.3.5.2 State 2. Envisioned

By contrast with the previous state, the artefacts can only be considered envisioned if there is a *specific vision* both for form and content. For example, for the language researchers' reports in Shakespeare Lives, this vision could be described as follows: a report should include a section on content analysis (based on Twitter and/or Seana Weibo data) and online ethnograpgy (based on observing Facebook/VK.com). Specifically for the content analysis section, it should go through the key human-coded variables, showing their distributions and displaying their relationships (thus not only *reflecting project goals*, but also providing findings in response to the *research questions*) and providing example tweets / Weibo posts to illustrate the key point. Such specified vision thus poses *requirements* for the artefacts – although these requirements cannot always be very precisely formulated. The detailed vision for artefacts can also include the vision for

ARTEFACTS: OUTLINED	*There is a tentative understanding of the possible forms of artefacts and the motivations behind them.*
Artefact types considered	There is an understanding of what types of artefacts may address the research goals and satisfy the stakeholders' needs.
Stakeholder input provided	At least some of the key stakeholders have explicitly specified what forms of artefacts they desire.
Outline complete	There is at least an understanding of the number and the types of suggested artefacts.
Research Goals: Formulated	*Potential goals of the social data science project are formulated.*
Stakeholders: Involved	*The stakeholder representatives are actively involved in the work and fulfilling their responsibilities.*

Table 4.21: Artefacts: Outlined. Conditions and prerequisites.

artefacts of interim use – e.g. in the Hit List production, it would specify the contents of chart spreadsheet prepared by the analyst for the production team (see Section 3.3.5.2). Finally, the envisioned artefacts should strive to *capture interests of multiple stakeholder groups*, including the underrepresented ones. For example, the sheet inclusion of VK.com into consideration while working on the Shakespeare Lives project was motivated by the desire to satisfy local branches of British Council – in this case, the British Council Russia in particular (see Section 3.2.4.1).

ARTEFACTS: ENVISIONED	*There is a clear, valid and agreed vision of the final artefacts.*
Vision specified	There is a clear vision of what should be the contents and the structure of each artefact.
Research goals reflected	It is clear how the vision of the artefacts corresponds with the research goals.
Research questions incorporated	it is clear how the findings derived in relation to the research questions can be incorporated into the artefacts.
Requirements formulated	What is required from the research artefacts is clearly specified.
Agreement achieved	The vision is agreed by the stakeholders and the team.
Research Questions: Outlined	*The research questions are formulated as a broad framework that sets the direction for the project.*
Stakeholders: In Agreement	*The stakeholder representatives are in agreement.*

Table 4.22: Artefacts: Envisioned. Conditions and prerequisites.

4.3.5.3 State 3. Supported

The "Supported" state is achieved when the social data science project as a whole can support the actual process of creating the artefacts. As such, the primary practical purpose of considering this state separately is in making sure that all the prerequisites of artefact creation, including the understanding of *how to create them*, are in place and that the *resources are sufficient* (see Table 4.23). The specific list of prerequisites is quite natural – understanding how to achieve the artefacts means that the work must be prepared, the resource base in terms of team, sources of data and infrastructure should be in sufficient state and the the understanding of how it all falls under ethical and legal compliance must be firm. In a way, the precise list of prerequisite alphas and their particular states may be flexible, but I believe it is important to have those points in the project when the team member is forced to remind themselves about all the aspects of the project together – and it is more dictated by the SEMAT's philosophy of kernel approach to project management (Jacobson et al., 2013) than necessarily particular episodes in the fieldwork.

ARTEFACTS: SUPPORTED	*The team is ready to engage in the research work that ultimaly results in the artefacts.*
Path to artefacts established	It is clear how the work planned for the project can lead to the artefacts.
Resources are sufficient	The available research resources (data sources, infrastructure and compliance) are sufficient for the planned work.
Team: Formed	*There is a sufficient number of engaged team members to progress towards the team's goals.*
Work: Prepared	*The conditions required for the work to start are met.*
Compliance: Strategised	*The team members share a clear understanding of how they can achieve ethical and legal compliance.*
Data Sources: Selected	*A firm decision on the use of each data source is made.*
Infrastructure: Demonstrable	*A prototype executable version of the infrastructure is available and demonstrates that the architecture is fit for purpose and supports testing.*

Table 4.23: Artefacts: Supported. Conditions and prerequisites.

4.3.5.4 State 4. Iterated

The fieldwork suggests that in a project with high degree of *stakeholder involvement* artefacts may get into a cycle of many iterations after the main analytical work is already done. In Shakespeare Lives and InfoMigrants projects, the *draft* reports that the teams had prepared got scrutinised by the British Council and the InfoMigrants editorial team respectively. They would provide *feedback*, question particular findings and suggest additional angles for their interpretation.

Prerequisites of this state may again seem quite self-explanatory in that a team is likely not to get to this state without *performing and being in control of the work* they have to complete. It is noteworthy that prerequisites include *execution of analysis methods* rather than answering research questions. There are two reasons for this: (a) research questions can hardly be considered answered until the interpretation of findings is agreed by all key involved parties – and this is a property of the next state of the "Artefacts" alpha and (b) not all analysis outcomes represented in artefacts necessarily answer any particular research question. For example, the outcomes on Strand 3 of the Shakespeare Lives were illustrative (and thus enriched the artefacts), but could hardly be considered findings by themselves (see Section 3.2.1.3).

ARTEFACTS: ITERATED	*The research methods are executed and inform an iteration of artefacts with the research findings.*
Relevant knowledge emerges	The executed methods produce knowledge that can be incorporated into the artefact drafts.
Drafts prepared	The team iterates through the draft versions of the artefacts.
Feedback received	Stakeholders provide continuous feedback on the drafts.
Progress being visible	Each iteration brings improvements to the artefacts.
Analysis Methods: Executed	*The methods are utilised for the needs of the project.*
Team: Performing	*The team is efficient and effective at progressing its work.*
Work: Under Control	*The work goes productively and in a risk-managed environment.*

Table 4.24: Artefacts: Iterated. Conditions and prerequisites.

4.3.5.5 State 5. Released

The last state of the "Artefacts" alpha is achieved when the finalised project artefacts are *made available* to the *project stakeholders* (see Table 4.25). The artefacts should *meet the requirements* – although, as the fieldwork suggests, in social data science the requirements are often quite broad and open; therefore, meeting them in practice means that the stakeholders are happy with a particular artefact iteration and do not provide more feedback. The finalised artefacts must incorporate the *answers to research questions*. Within the studied projects, the answers were included in numerous ways – from being postulated as findings and recommendations in the Shakespeare Lives project reports to being presented as the Hit List show on radio. Beyond the fieldwork, a crucial form of incorporating answers into project artefacts takes a form of decision-support systems and applications, with recommender systems being one of the most common examples.

ARTEFACTS: RELEASED	*Research artefacts are produced and presented to the stakeholders.*
Appropriate artefacts produced	The team has produced the artefacts in accordance with the agreed vision.
Artefact requirements met	The requirements formulated for the artefacts are met.
Answers incorporated	The artefacts capture the answers to the research questions.
Availability ensured	The artefacts are available to all agreed parties.
Research Questions: Answered	*Through applying the analytical methods to the acquired- and quality assured data, actionable answers to the research questions are derived.*
Stakeholders: Satisfied with Progress	*The stakeholder representatives see the progress towards their needs.*

Table 4.25: Artefacts: Released. Conditions and prerequisites.

4.3.6 Compliance

The "Compliance" alpha was informed by the experience of all case study, including (and predominantly by) that of the project on studying the Dark Net criminal markets (see Section 3.4.2). As already mentioned, this project was also used for the internal evaluation of the Scorecard Deck (see Section 5.2). To keep evaluation as separate from gathering primary evidence as possible, no drafts of the "Compliance" scorecards were presented in the Dark Net project. To compensate for the omission of formative internal evaluation for this alpha, I have heavily grounded the scorecards in literature.

4.3.6.1 State 1. Considered

The first state is achieved when the ethical and compliance issues associated with a social data science project and informed by *its goals* are identified (see Table 4.26). The types of *research risks* are derived from Martin and Christin (2016), who identifies groups of risks associated with various parties involved with the project. The Dark Net project team used this paper to distinguish key areas of concern that had to be dealt with. Some of those risks had *legal implications*, as accessing data from Dark Net implied needing to take precautions that no of them would contain materials possession of which is illegal by itself (e.g. images of child abuse). But even when dealing with legal data, there still are applicable legal frameworks. For example, the General Data Protection Regulation (GDPR) limits researchers' rights to store their data on European residents on servers outside Europe (European Parliament, 2016).

Compliance with the terms and conditions of the selected *data source* and *elements of the*

technical infrastructure was relevant for most studied projects. Indeed, the online platforms whose data were employed in the projects all had policies restricting use of acquired data – e.g. the Twitter Developer Policy (Twitter, 2017) and the Facebook Platform Policy (Facebook, 2018b). The same can be said about the employed data analysis tools. For example, Sysomos sets limits on dissemination of data visualisations that it provides to "limited and insignificant excerpts" (Sysomos, 2015) – although in practice these limitations did not affect the work on the Shakespeare Lives evaluation, as the project only employed Sysomos for data acquisition (see Section 3.2.6.2). As will be shown further (see Section 4.3.6.3), practical implementation of terms and conditions of data sources and technologies is often a compromise (Voss et al., 2016).

Finally, as shown in Section 2.3, the *public* discussion of ethics and compliance in social data science is ever evolving, and a researcher could be advised to follow it not only to comply with public expectations, but possibly also to discover ethical implications of their work that they have failed to think about themselves.

4.3.6.2 State 2. Guided.

To progress a project's compliance to this state, the team members should inform themselves on the relevant best practices, guidelines and advice that should be applicable to their project given the *scope of its compliance issues* (see Table 4.27). The guidance may be *professional*, such as the currently developed IEEE standard on Algorithmic Bias Considerations (IEEE, 2017) or *governmental or intergovernmental*, such the Universal ethical code for scientists in the UK (Department for Business, Energy & Industrial Strategy, 2017). The guidance may also come from the *organisation(s) hosting the team* and from the *project stakeholders*. For example, in the Dark Net project, even though the team members were thorough in their consideration of the compliance issues on their own, knowledge of the institutional requirements could have allowed for a quicker project start (see Section 3.4.2.4). Besides, in one of his earlier projects, the PI of the Dark Net study employed good compliance practices suggested another researcher team in the field, Barratt (2011) (see Bancroft and Reid, 2016).

4.3.6.3 State 3. Strategised

Strategising compliance in a social data science project involves elaborating on concrete measures that have to be in the project (see Table 4.28). In the Dark Net study, this required balancing complex interests of explicit and tacit stakeholders who represented *different groups of society* – from law enforcement to cryptomarket participants (see Section 3.4.2.2). Additionally, we had to develop *risk management procedures* to protect ourselves and out project infrastructure from possible reactions from the marketplaces in response to our scraping attempts (see Section

COMPLIANCE: CONSIDERED	*It is identified which ethical and legal concerns the project raises.*
Research risks identified	There is an initial understanding of how achieving research goals may cause harm for the studied groups and individuals, for the research team and for the key stakeholder groups.
Data sources compliance considered	For the identified potential data sources, there is an understanding of the conditions on which the researchers may acquire, use and share the data, and disseminate data products.
Technologies compliance considered	For the identified candidate technologies, there is an understanding of the conditions on which the team can deploy them, share access to them and disseminate products of their deployment.
Legal requirements considered	Common law context relevant to social data science projects and to the specific topic (e.g. data protection acts) is considered.
Current discourse considered	The team is familiar with the recent ethical and other compliance controversies, scandals and generally 'hot topics' in the fields related to their proposed research topic and research methodology.
Research Goals: Formulated	*Potential goals of the social data science project are formulated.*
Infrastructure: Architecture Selected	*Architecture has been selected. It addresses the key technical risks and any applicable organisational constraints.*
Data Sources: Identified	*The possible relevant data sources are identified and a preliminary list of appropriate data sources is compiled.*

Table 4.26: Compliance: Considered. Conditions and prerequisites.

3.4.2.3).

As can be seen, the developed policies required strict, disciplined following on behalf of the team members. That could not be achieved without a *shared understanding* of them. To achieve it, our team had many meetings specifically on the compliance issues. Those meetings allowed us to both discuss these issues conceptually and coach each other on the particularities of implementing them. We wrote down the instructions or using the *emerging compliance infrastructure* and went through several iterations of ethical statements, thus forming a *corpus of compliance statements*.

As mentioned above, *compliance with license agreements* was a concern in all studied projects. However, the project teams often made a decision to follow the spirit of those license agreements rather than the letter, especially when considering terms of conditions of online platforms that were used as data sources. The license terms were arguably often written with particular use of

COMPLIANCE: GUIDED	*The team has the required guidance to formulate a compliance strategy.*
Scope of the issues understood	The team has an understanding of the severity of the identified compliance risks and the chances of some risks being unidentified.
Own experience considered	The team members draw on their past experience of dealing with similar compliance issues.
Other experience considered	The team members have accessed (through such channels as personal connections or reading literature) relevant experiences of other teams in their field.
Organisational guidance studied	The team and its individual members have studied and understood the organisaitonal guidance of their respective hosting organisations.
Stakeholder guidelines studied	The team has studied and understood the available guidance provided by the key stakeholders (e.g. ethical guidelines by funders).
Professional guidance studied	The team has studied and understood available professional guidance such as industry standards or publications of pro-ethics advocate organisaitons.
Governmental guidance studied	The team has studied and understood non-binding guidance suggested by the state and by intergovernmental organisations (e.g. the Data Science Framework in the UK).
Stakeholders: Involved	*The stakeholder representatives are actively involved in the work and fulfilling their responsibilities.*

Table 4.27: Compliance: Guided. Conditions and prerequisites.

data in mind. For example, the Twitter Developer Policy (2017) forbids any transformations of content received via any Twitter API, including *translation*. For the Shakespeare Lives project, where the Twitter data analysts formed a multilingual team with each researcher focusing on a particular language, this term effectively forbade the researchers to translate the user biographies that had been collected with the Twitter REST API (see Section 3.2.6.1). If the team had adhered to the Policy, some of the qualitative findings could have not been shared neither within the team nor with the stakeholders. Yet, this term rather seems to be aimed at developers of web applications that who display Twitter content to their users, as Twitter may be interested in preserving the original look of its content. Therefore, the Shakespeare Lives team members did not abstain from translating a user bio when required; however, they took a screenshot of its original version and put a textual translation nearby.

Following the license terms is often also practically impossible. The same Twitter Developer Policy prescribes to delete data if they represent content that has been deleted or otherwise taken

out of public access on Twitter "as soon as reasonably possible". Within the Hit List production this effectively meant "never". Indeed, the tweets were collected in real time through the Twitter Streaming API as part of 1% sample (see Section 3.3.3.1). Therefore, the tweets that would later be deleted or by users whose accounts would later be suspended had equal chances to be acquired as those tweets that would remain intact forever. Tracking the fate of the individual acquired tweets would have meant sending repeated requests to the Twitter REST API that had been empirically found severely rate-limited. Hence, the only measure that was taken in Hit List production to comply with this policy element was to only use individual tweets for usual modes of analysis that based on aggregation of data. If the Hit List presented was ever to read a tweet during the show and if the tweet was publicly unavailable at the moment, she would make such a caveat.

COMPLIANCE: STRATEGISED	*The team members share a clear understanding of how they can achieve ethical and legal compliance.*
Guidance internalised	The team has critically assessed the available guidance and adapted it to their views and to the project's curcumstances.
Legal and license requirements factored in	The team has worked out how the legal requirements and license conditions can be met within the project.
Society interests factored in	The team has made sure that the project is sufficiently aligned with the wider society interests.
Need in consent evaluated	There is understanding of whether the project will at any moment require explicit consent to participate from the studied individuals.
Understanding shared	The team is not only in a principle agreement over the project compliance, but actually shares understanding of how to deal with specific compliance issues.
Compliance statements formulated	All the required compliance statements such as the ethical statements and the internal compliance policies are formulated, strongly preferably in writing.

Table 4.28: Compliance: Strategised. Condition. No prerequisites specified.

4.3.6.4 State 4. Progressed

This state of the "Compliance" alpha effectively represents taking initial actions on all the decisions made while developing compliance strategy (see Table 4.29). If a project team has formulated an ethical statement, the team has to *submit this statement* for ethical review. If they have designed a compliance infrastructure, they have to *construct this infrastructure*. In practice surely progression towards these two states often happens simultaneously. As mentioned above (see Section 4.3.6.3), the Dark Net project team put the infrastructure and the models of its use

together simultaneously. However, conceptually "Strategised" and "Progressed" are two distinct states, and it is valuable to remind researchers who use the Scorecard Deck to act upon their decisions. Additionally, the to achieve this state a research team may have to actually *adjust their activities* and *align research practices* with the developed compliance strategy. For example, in the Dark Net project we decided to use Tor to scrape clear net website data, even though it was not technically necessary, to protect our identities (see Section 3.4.2.3).

COMPLIANCE: PROGRESSED	*The team has taken the steps necessary to secure the compliance resources and is ready to actively engage in the main bulk of research subject to secured ethical clearances.*
Compliance infrastructure in preparation	The team has started to put up the components of the infrastructure that would facilitate compliance in data acquisition and analysis.
Means to assure consent prepared	Information sheets, debriefing practices, consent forms, participant agreements and other means to seek and secure participant's consent are ready for use.
Clearance applications submitted	Applications for required clearances, such as the ethical statements or the applications for accessing highly sensitive data, submitted.
Planned activities adjusted	The planned research activities are re-validated and adjusted to be compliant with the established compliance strategy.
Research process being aligned	The team is capable of a fair judgement of when each of the exisiting compliance issues can be resolved to which degree. The team can align the research process with this judgement, so that it always strives to be compliant.
Infrastructure: Demonstrable	*A prototype executable version of the infrastructure is available and demonstrates that the architecture is fit for purpose and supports testing.*

Table 4.29: Compliance: Progressed. Conditions and prerequisites.

4.3.6.5 State 5. Secured

This state represents achievement of results of the compliance-seeking activities designed and performed within the two previous states of the alpha (see Table 4.30). As such, its conditions are supported by the same fieldwork evidence.

4.3.6.6 State 6. Maintained

The final state of the "Compliance" alpha is achieved when those factors that have motivated compliance-seeking activities in the first place (see Section 4.3.6.1) are monitored and the changes to them are incorporated into the compliance strategy and implementation (see Table

COMPLIANCE: SECURED	*The compliance resources required for the project are secured to the degree that the team can fully engage in the main body of the research work.*
Ethics cleared	If relevant, the team has received ethical permissions to carry the research from the hosting and funding organisation and other relevant stakeholders.
Consent sought	The practices aided at seeking participant consent are continuously and consistently implemented.
Access permissions cleared	If the project involves access to data that are highly sensitive (e.g. medical records), legal clearances for accessing those are obtained.
Agreements achieved	For each data provider and infrastructure component vendor, there is an agreement in place about using their provided services within the project and disseminating the outcomes of such use.
Compliance infrastructure set	There is a working infrastructure that supports carrying out research in compliance with the ethical and legal requirements.

Table 4.30: Compliance: Secured. Condition. No prerequisites specified.

4.31). This state is mostly informed by the same background research that has informed Section 4.3.6.1. If the public discourse on compliance and ethics of doing social data-driven research is continuously morphing and the legal and license requirements are adjusting accordingly – the change to terms of use of Facebook Graph API because of the Cambridge Analytica scandal is an example (see Section 2.1.2) – being compliant with the norms at the moment does not automatically imply maintaining compliance consistently.

At the moment of writing, it seems like many changes in discourse on ethics and their practical implications could have been *forecasted* and appropriate *responses* could have been developed. However, this may be hindsight bias (Christensen-Szalanski and Willham, 1991). It is still undeniable though that some other changes to compliance norms *can* be anticipated. For example, the GDPR European Parliament (2016) came into effect only in 2018, almost 2 years after it had been developed and published. Therefore, any research team whose data treatment were affected by GDPR should have had enough time to prepare for the changes.

COMPLIANCE: MAINTAINED	*The team works on the project while reactively and proactively maintaining compliance.*
Team in compliant work	The team is engaged in the main bulk of the research and does it in a compliant manner.
Changes to project setup monitored	There is continuous monitoring and compliance assessment of the ongoing changes to the team composition, the work involved, the data sources, the infrastructure used and other project setup characteristics.
Changes to compliance resources monitored	There is continuous monitoring and compliance assessment of the ongoing changes to the governing agreements, permissions, clearances, laws and other compliance resources.
Changes to wider compliance context monitored	There is continuous monitoring and compliance assessment of the ongoing changes to the public discourse on ethics and compliance, ethical scandals and other wider compliance context.
Potential further changes forecasted	The knowledge of the compliance resources and of the wider compliance context is sufficient to forecast potential changes to the formal and informal compliance requirements.
Responses developed	There are measures in place to assure that if the circumstances do change, the compliance procedures will be adjusted accordingly.
Team being confident	Team is confident that they can continuously stay in compliance.

Table 4.31: Compliance: Maintained. Condition. No prerequisites specified.

CHAPTER FIVE

EVALUATION OF SOCIAL DATA SCIENCE SCORECARD DECK

In the previous chapter I provided an overview of the Social Data Science Scorecard Deck and linked its contents to the fieldwork discussed in Chapter 3. In this chapter, I would like to report on the evaluation of the Scorecard Deck, which constitutes the last piece of research undertaken for this thesis.

Borrowing the terminology from the field of performance measurement in education, the evaluation seeks both *summative* and *formative* assessment (Wholey, 1996; Taras, 2008) of the Scorecard Deck. As discussed in Section 3.1, the process of informing a tool with ethnography is often met with a range of certain criticisms. Summative feedback would allow to see whether the associated limitations have been surpassed so that the scorecards actually generalise to other social data science projects and are capable of adding value to those. Formative feedback, in turn, would play an even more important role of *helping to make the Scorecard Deck better*. The Scorecard Deck provides scope for future iterations – its content can be easily expanded by addition of new alphas and revision of their states and conditions. There are multiple strands of further work on the Scorecard Deck that lie beyond the scope of this thesis (see Section 6.1). The evaluation can provide direction to such further developments. Given these considerations, I can proceed to introduce and justify the chosen evaluation methodology.

5.1 Methodology

Conducting an evaluation of the Social Data Science Deck requires paying attention to several specific features of the subject matter. By its very nature, the Scorecard Deck is a multi-component tool that is designed to provide long-term benefits for social data science projects. Those benefits may be, amongst others, a shared understanding in an interdisciplinary team, increased research quality, ethical and responsible research conduct and greater consideration of needs of various stakeholder groups including the tacit ones. All of these are problematic constructs that are not always directly observable. An attempt to reduce them to a set of measurable variables and to generally base the evaluation on quantitative measurements would therefore inevitably lead to low *construct validity* (Smith, 2005) – the findings would not represent the phenomena of interest. It is a lot more productive to instead ask the questions of *how* the potential benefits of the Scorecard Deck may be manifest in practice, *how* the specific context of a social data science project may affect the scorecard's value, and *how* the scorecards can be further improved. All this calls for *qualitative research methods*.

Furthermore, an attempt to evaluate the Scorecard Deck in a highly controlled environment, e.g. using some predefined scenario for how a social data science project unfolds will lead to low *external validity* (Siegmund et al., 2015), i.e. a lack of transferability of findings to often very different settings (Polit and Beck, 2010). First, using a standardised scenario would understate the variety of social data science project. Second, even if a *set of different scenarios* is employed instead, those would bear a high risk of being too schematic. Uncertainties are a fundamental component of a research process (Grabe, 2005) and it is hard to represent those in an artificially constructed – and reasonably (for the purposes of a controlled evaluation) short – scenario. The evaluation should instead seek to apply the scorecards in the context of *naturally evolving real-world social data science projects*.

The considerations above suggest that a traditional experimental evaluation design, despite their strong internal validity would not be appropriate for the task at hand (Lazar, 2010). Instead, arguably the most robust approach to evaluating the Deck would *in principle* be employing a number of in-depth case studies with each case being a thoroughly studied social data science project that is managed with the help of the Deck. If the number of the cases is sufficiently large and their settings are sufficiently varied, the findings can have strong construct validity as they would lead to deep understanding of the Deck's merits and limitations while retaining both internal validity through triangulation of results received with different methods and from different setting (e.g. team meetings versus individual use) and external validity and generalisability through cross-case synthesis of findings (Yin, 2013; Diefenbach, 2008).

That said, practical feasibility of such an approach is questionable, as social data science projects often take months to years. Moreover, in the face of a very unlikely, but still existing chance the Scorecard Deck could turn out to be *detrimental* to social data science projects, it seemed irresponsible to suggest to use it in a large number of real-world research projects until its value is somehow validated. As a result, I opted instead to use two different evaluation modes that are each feasible to execute and, when used in conjunction, harness most of the benefits of the multiple case study approach.

The first round of evaluation was *depth-oriented* and consisted of **a single case study** (see Section 5.2). Similarly to the case studies discussed earlier in this thesis, I took part in this project as a **participant observer**. During my three months in this capacity, our team used the Scorecard Deck for different purposes (to clarify the project goals, to identify key gaps and to track progress) and in different settings (in a team meeting, solely by the principle investigator and solely by myself). This allowed me to triangulate observed effects of the scorecards in various contexts and also to capture the how the use of the Deck was evolving as the project was progressing and as the team members were getting more experienced with the tool. I was also able to clarify and validate the findings through a final interview with the principal investigator. The feedback gathered through the case study demonstrated the value of the Scorecard Deck in a project that was significantly different in its goals and set-up from the ones used to inform the Deck. The findings also suggested a range of improvements to the tool that were implemented in-between the evaluation rounds.

The second round of evaluation was *breadth-oriented* and consisted of **a series of nine interviews with experts in social data science**. This round compensates for the limitations of the single case approach. It shows the Deck's applicability to- and potential value for social data science projects that take place in a wider variety of settings. During the interviews, the respondents were asked to retrospectively apply the scorecards for one or more alphas to one of their past social data science projects, or to generally talk about the aspects of one of their past projects that were covered by the Deck. In that, the interviews, while sharing some structural similarities, were highly reactive to an interviewee's responses and as such fitted the bill of being an adaptive qualitative research method. This allowed me to harness the value of "probes" (Boyce and Neale, 2006, p. 5) – questions that ask a respondent to reflect more deeply and critically on a previously raised point. It also allowed me to use the experience of past interviews to inform subsequent ones. Such continuous question refinement, while it may be seen as a methodological flaw within quantitative evaluation methods, becomes "a sign for progress [towards] an increasingly better and deeper knowledge" in qualitative data collection (Diefenbach, 2008, p. 877). Finally, the approach taken allowed to make use of *real-world*

scenarios of social data science projects that stemmed from the respondents themselves and thus did not suffer from sketchiness and oversimplification.

Now that the two evaluation modes are introduced and their complementary nature is discussed, I will proceed to discuss each of those, starting chronologically with the in-depth evaluation.

5.2 In-depth Evaluation: Applying the Deck in a Social Data Science Project

The first round of the Scorecard Deck evaluation was performed through studying the use of this tool for managing a social data science project studying the resilience of cyber-criminal marketplaces. Some aspects of this project have already been covered in Section 3.4.2. In accordance with the terminology introduced in that section, I will further call it *the Dark Net study* (or *the Dark Net project*. The project was hosted in the University of Edinburgh and was led by Dr Bancroft – a social scientist with a long-standing research interest in the social life of substance users (cf. Bancroft, 2009) and a more recent interest in online trade in illicit goods (in particular, but not exclusively, drugs, cf. Bancroft and Reid, 2016). Besides Dr Bancroft, the project team included Dr Alex Voss (the supervisor of this thesis) and myself from the computer science side.

The project had been funded by an ESRC *Impact Grant* several months before I joined the research team. According to the grant application form, the project's primary objectives were to study how cyber-criminal marketplaces react to law enforcement interventions (e.g. marketplace shutdowns or major vendor arrests) and other disruptions (e.g. large-scale scams), to elicit the social and technical factors that allow such marketplaces to be resilient to the interventions (i.e. to quickly restore the total volume and frequency of trade across the popular marketplaces) and to ultimately aid law enforcement in prioritisation of their investigations. The work on the project was planned to be carried out in close cooperation with a particular partnering law enforcement body that provided a letter of support for the grant application with a promise to aid the project both in cash through financing the research activities and in kind through investing time of one the body's employees.

Unfortunately, the impact partner failed to follow up on their initial promise of support. As a result, even though some pilot development work on the tools for acquiring data from the Dark Net was carried out, the start of the *substantive* research work had to be shifted by the project investigators several times in anticipation of resuming contact with the law enforcement body. At a certain moment of time, the pressure of the reporting deadlines imposed by the Impact Grant motivated the investigators to initiate the research in absence of the partner's involvement. At that time, among other things, the investigators included me into the team as a part-time research assistant and, as I had had a chance to briefly advertise the idea of scorecards to Dr Bancroft and to attract his interest, agreed to attempt using the Scorecard Deck to manage the project. The challenging project circumstances made the Dark Net study an excellent case for evaluation of the Social Data Science Scorecard Deck. Indeed, they allowed us to use the Scorecard Deck not

only in tracking the ongoing progress, but also in setting up the work. Our team had to effectively redesign the study so that we could progress towards impact even if the partner organisation would not join us, while also being constrained by time scales and budget.

Another factor that contributes to the value of the Dark Net project as an evaluation case study is that it significantly differed in its key characteristics from the projects employed to inform the Scorecard Deck. The relative flexibility of the Dark Net projects' goals can be juxtaposed with quite a clear direction that the previous projects enjoyed: in those, the work had to contribute to the creation of specific artefacts (be that multiple evaluation reports or a weekly radio show) and to address the needs of specific key stakeholders (be that consultancy customers or a broadcasting company). The subject matter by itself is a substantive difference: studying evidence of criminal activities raises many specific issues, some of which have already been covered in Section 3.4.2. The nature of the data and what interactions they represented were also different: the other projects mostly focused on social media data that reflected user interactions with some online content. The Dark Net project, in contrast, focused on marketplace data (such as transaction counts and reviews) that reflected the markets' buyer interactions with particular buyers and products. The Dark Net project also employed forum data that represented extended, tightly-connected conversations. Given all the above, *if* the Scorecard Deck was to be found an effective tool in the Dark Net case study, that would *strongly suggest that the scorecards do not overfit to the projects that informed their design*.

NB: before starting the discussion, I would like to remind readers that the scorecards were refined after the Dark Net study. The interesting major changes will be discussed in detail, while the minor and trivial ones such as changes to wording or splitting/merging consecutive states within an alpha may be left without a mention for brevity and narrative consistency. Either way, readers should not be surprised if they encounter a mention of an alpha, state or condition that do not entirely match the contents of the Deck presented in Section 4.3 and Appendix A.

5.2.1 Shaping the Research Goals

Even though throughout the course of the Dark Net case study our team systematically worked through the scorecards for most alphas, it was the "Research Goals" alpha whose associated scorecards provoked the most extensive discussions and to which we returned most often. The scorecards associated with this alpha were initially worked through at the first team meeting that I participated in. At that time, I had only a vague understanding of the research area in general and of the project state. I had an impression that its overall design and goals had been already firmly set and expected the investigators to quickly talk through the scorecards for several key alphas and thus introduce me to the project. Counter to my expectations, the discussion that was

inspired by the "Research Goals" scorecards took almost an hour and a half of that team meeting and involved making several crucial research decisions.

5.2.1.1 Research problem and stakeholder interests

As the first scorecard – "Research Goals: Idenfied" – suggested, we started the discussion by spelling out the problem that the research aimed to solve. Dr Bancroft suggested:

> "I think the problem and is with that [...law enforcement bodies] can assess the financial size of different operations [of the cybercriminal marketplaces], but that doesn't really tell them much about the actual social dynamics [...such as] trust relationships [that shape the] responses to law enforcement intervention."

As can be seen, this perspective closely matches that expressed in the Impact Grant proposal. It is focused on the interests of the law enforcement partner and it suggests to solve specifically *their problem*. As will be shown, the subsequent discussion did not completely change this perspective, but it *shaped and enhanced* it. Another condition of the "Research Goals: Identified" scorecard – identification of other stakeholder groups – came to play an important role in that. Dr Voss contributed the first suggestion:

> "[W]e have some awareness of other stakeholders, which is the market participants. You know, those are the ones I would [...] list for now."

Arguably, we (or at least Dr Bancroft and Dr Voss) had indeed implicitly recognised the market participants as a stakeholder group and had acknowledged their interests at the back of our minds. However, only after pronouncing this seemingly trivial idea explicitly and in the context of working with the "Research Goals" scorecards, that we, as a team, actually started to consider the interests of market participants as a *major factor underpinning our research plans*. This is strongly indicated by the suggestions by Dr Bancroft and Dr Voss that followed almost immediately:

> "I think we'd also want to focus on reducing harm. So our findings will be about *reducing harm*, rather than necessarily *reducing the operation* of the markets as such. [...Law enforcement bodies] want evidence of the effect of their operation, so what we're doing *isn't going to guide* their operations." (Dr Bancroft)
>
> "There's a recognition that the law enforcement have been good at disrupting, [but] less good at assessing the impact of disruption. [...] The impact on disruption [and]

> the impact on reduction [of harm are] two different things. [...] [Law enforcement] are very good at doing one thing, but may be unsure whether they're actually achieving the latter." (Dr Voss)

One may notice that the latter quote resembles the earlier statement of the research problem, but with an important change in focus: rather than casting the problem as the law enforcement's difficulties in preventing online trade, we cast it as difficulties in *reducing harm*.

It is worth noting that harm reduction, at least under the assumption of "good policing", is *an explicit goal* of law enforcement (Ratcliffe, 2015). Hence, this shift of focus did not suggest abandoning the interests of the the impact partner whom Dr Bancroft and Dr Voss still expected to resume contact with at that time. Moreover, it did not imply abandoning the objective of modelling the social relationships exhibited by cryptomarket participants. For example, later in the meeting, when discussing the "Research questions outlined" condition of the "Research Goals: Evaluated" state, Dr Voss brought in the issue of operationalising the trust relationships:

> "We certainly have to find ways of describing trust relationships and how they manifest themselves. So, I think that the link between trust and observable behaviour will certainly be part of [...] the research questions." (Dr Voss)

That said, the shift of research problem interpretation towards harm reduction had several interrelated consequences. The first one was in strengthening the common ground and shared understanding of the proposed research. The ethos of harm reduction was something everyone of us could (and was happy to) subscribe to. Moreover, Dr Bancroft's track record of research into communities of drug users strongly hints at a chance that our research would have eventually converged on harm reduction even if we had not had this project management session with the scorecards. However, even if this is the case, establishing the focus on harm reduction without the help of the scorecards would arguably have taken a longer time and it would possibly never have been made *explicit*. Therefore, a lack of common understanding might have hindered progress and project success. Incidentally, Dr Bancroft expressed a similar point of view in the interview at the end of the case study:

> "I think it would be a lot more difficult to manage [this project] without this [tool]. [...] In my experience of any kind of team – even ones where [the members are from] basically the same discipline but have got different elements to work on – it's really hard to coordinate [...] across the different parts of it. [...] It would be more

> difficult [...to understand] why *particular priorities are being set.* [...] I think having something a bit more structured definitely helped."

Recognition of harm reduction as a priority early on in the team meeting allowed us to broaden the scope of what we could consider good research outcomes. For example, Dr Bancroft suggested that we could help law enforcement with modelling trust in cyber-criminal communities in such a manner that would allow law enforcement to "use trust to transmit harm reduction information to people visiting the markets [...] early on [in their visits]". By that, he outlined the possibility for our research to support *non-disruptive* law enforcement interventions. This line of discussion reached its conclusion when we, in correspondence with the suggestions of the "Research Goals: Evaluated" scorecard, attempted to formulate what the potential societal value of the Dark Net project could be. We decided that the best-case achievement of our work would be a contribution to a *shared agreement* among different stakeholders on the priorities in harm reduction activities and, consequently, in interventions against cybercriminal marketplaces. In accordance with the harm reduction ethos, such an agreement could potentially be useful for many stakeholder groups – even the ones who could not be included in it:

> "I think, [...] realistically the agreement would be between [...] law enforcement and the government, or between different parts of law enforcement. But the benefit would also be for market participants. [...] It might give some guidance [...] to platform operators as to [...] what to do not to get into the focus of law enforcement. [...] They might change the rules under which their markets operate to avoid [...] the complete take-down." (Dr Voss)

5.2.1.2 Public-facing impact

As far as focus on harm reduction broadened our the research goals and desired research outcomes, it could also suggest a broader spectrum of ways to achieve impact. Given the uncertainty regarding the impact partner's involvement, identification of alternative impactful activities was a priority for the Dark Net project. In the discussed team meeting, one idea on the alternative pathway to impact initially stemmed from the continued discussion of the stakeholder groups. Specifically, Dr Voss observed that the society as a whole were also a stakeholder in the Dark Net project:

> "Society in general has got the interest in [...] good policing and reduction of harm. And [since] we would claim that this is an impactful project, [...] it's worth [considering] this."

Although by itself this observation did not seem to require specific actions, the subsequent discussion of the scorecards provoked some actionable suggestions. One condition of the "Research Goals: Demanded" state required us to envision research artefacts that could be of value for the stakeholders. At this moment, Dr Voss followed up on his prior observation by suggesting to produce "an impact report, which is different [...] from [both] the academic project report and the report for law enforcement". Dr Voss suggested to use such an impact report to kick-start dissemination of findings to the general public and Dr Bancroft supported this suggestion:

> "I'm thinking about [...making an] appearance [in regional media], where we [could] talk about the project [and] tell the general public [...] what great things we have achieved that make them safer. So, I think the impact report is not just a report [–] it should actually be a, well, kind of working document that prepares us [...] to achieve and measure impact." (Dr Voss)
>
> "We need an impact case study, so we need some kind of documents to trace impact." (Dr Bancroft)

Even though no particular details on how the impact report should look were discussed, Dr Voss felt that the idea was valuable by itself:

> "Having that kind of thought planted earlier, I think, will then help us to [...] generate impact and be prepared for [dissemination activities]."

5.2.1.3 Reflecting on the research goals discussion

Before moving on to the analysis of scorecard use in other environments, I would like to reflect on the overall experience of shaping research goals with their help.

First of all, it seems that the scorecards provoked the biggest sparks of creativity when the discussed conditions were interlinking different alphas: even though we talked through the "Research Goals" set of cards, the crucial conditions asked about artefacts and stakeholders. One explanation for why such conditions were so effective is that they allowed to bring different project aspects together and truly think about the project holistically. This is an important notion, since some of the later observations showed the flip side of such conditions: annoying redundancy (see Section 5.2.3.1). The need to balance these two concerns ultimately led to introduction of the concept of *prerequisites*. Most cards in the current version of the Scorecard Deck (see Section 4.3) have separate prerequisites and conditions.

Second, it is interesting how gentle the learning curve of using the scorecards was, considering that for both investigators it was the first time they actually used the tool. For example, the first crucial suggestions about the stakeholders were made about 15 minutes into the meeting. This observation is reflected by Dr Bancroft in the final interview:

> "I remember thinking back when I looked at [the Scorecard Deck], I got an instant sense for what it is about. [...] It was very user-friendly from my perspective."

This does not mean that the scorecards require no introduction – at the end of the day, I was actively involved in that first session and could thus steer the discussion when the investigators showed hesitation. Yet, as the discussion progressed, this effectively became unnecessary. Again, this observation is in line with Dr Bancroft's assessment:

> "I think, [...] you would need at least a session to go over, in some form, what [the process of using the scorecards is] about, [...] which is totally acceptable. [...] Now I would have had confidence in using [the scorecards on my own]."

Third, in relation to the idea of producing an impact report, I would argue that the sheer process of using the scorecards would be immensely useful. As Dr Bancroft pointed out in the interview:

> "Possibly the most enticing thing [about the scorecards] for me is [using them to] translate qualitative research into a form that can be easily transmitted to different audiences. [...] There's a certain kind of subjective vagueness in a lot of qualitative research. [...] I think that [...] a real problem [is] in how we translate and communicate that there is actually some rigorous process. [...] Scorecards can be very useful in that because they break down what's often a[n involved] process into specific achievable points. [...Scorecards] can be used particularly to communicate at the end of the project [...] what has actually being done there."

To further emphasise the value of scorecards for impact reporting, I would argue that the process of using the scorecards in their spreadsheet implementation produced an important project artefact by itself. Throughout the Dark Net project, we did not only use the scorecard spreadsheets to assign progress scores to the conditions, but also provided interpretations for those conditions and indicated what specifically had been achieved and what required further work. A quick skim through these spreadsheets could therefore provide a reasonable overview of the project state. Figure 5.1 shows an example spreadsheet from the project's scorecard deck.

5.2.2 Assessing Project State

The next session of using the scorecards in the Dark Net project was focused on assessment of the project's overall state by working through several alphas. In this exercises, Dr Bancroft effectively worked through the suggested scorecards alone: Dr Voss was absent in that meeting, and I deliberately took an observational stance. This does not mean that I kept silent throughout the process: if Dr Bancroft asked me to clarify the formulations of certain conditions or questioned the logic of the scorecards, I discussed those issues with him. However, I restrained myself from directing Dr Bancroft in his choices of alphas and states to evaluate and in his assessment of project's progress towards particular conditions.

Within a one hour session, Dr Bancroft easily managed to revise the scorecards for the "Research Goals" alpha and to go through the scorecards for the "Research Questions", "Data" and "Research Methods" alphas for the first time. Not only did he assess as many conditions as was reasonable at that stage of project progression, but also talked through his thinking process, so that I could then summarise it as notes. Figure 5.1 displays a screenshot of the filled-in scorecards for two states of the "Research Methods" alpha and the marginal notes.

The commentary that Dr Bancroft provided to the scorecards was quite detailed and demonstrated critical examination of the project circumstances. This suggests that even when used individually by a project manager rather than as a discussion point in a team meeting, the scorecards can foster reflexive thinking. For example, when revising the "Research Goals" alpha, Dr Bancroft offered the following detailed account of risks associated with the project and enhanced it with acknowledgement of potential countermeasures:

> "Primarily, the risks [are] interruption the data collection. [...] I think [that's a] major risk of a darknet market being shut down. [...] If this happens what we have to do probably either switch to a different market, or [...] go to the kind of secondary evaluation by looking at the data [that is already scraped by] somebody else [and] applying [...] our methods to that [data]. So, we would be looking more at the kind of resilience modelling aspect of it. So, so there is a risk there, but it's manageable."

The ease with which Dr Bancroft used the tool suggests that the logic of the Scorecard Deck fitted his own conceptualisations of the research process. The feedback that Dr Bancroft provided at the completion of the exercise confirms this and indicates that Dr Bancroft acknowledges the value of the Deck as specifically a *holistic* management tool:

> "[The tool] nicely fits the kind of logic model that I use for designing research.

Research Methods 1 / 4			
IDENTIFIED	*A tentative list of candidate research methods is compiled.*	**6 of 6 conditions met.**	
- Broad research questions considered	The broad research questions that reflect the research goals inform what kinds of methods may be used.	**Yes**	See Research Goals and Research Questions
- Prospective data types accounted for	The types of data relevant to the endeavour further inform the methods.	**Yes**	Yes, we have the forum data and the market transaction data. Edit 30/04/2018: more potential data types to come. Helper resources (reddit, deepdotweb).
- Traditional methods considered	Traditional research methods that can complement the social data science methods are considered.	**Yes**	Yes, and discarded within the project: out of scope + less obvious ways to impact. May be used in conjucting study.
- Triangulation considered	The methods allow cross-validating the findings derived with each other.	**Yes**	We've considered the relationship between qual and quant and how they complement/triangulate each other.
- Complementation considered	The methods allow studying the research problem from various angles.	**Yes**	Same as above.
- Preliminary list of methods compiled	A preliminary list of methods is compiled.	**Yes**	
Research Methods 2 / 4			
EVALUATED	*The merits and limitations of each source and of their combination are assessed to formulate a firm selection on of data sources.*	**4 of 5 conditions met.**	
- Mapping to research questions established	It is identified which methods are better suited to answering each specific research question.	**Yes**	Resiliance seems to be more answerable with quant market data with supplementary quals involved to triangulate. Trust seems to have qual focus.
- Validity assessed	It is assessed whether each method allows to progress towards the metrics that are valid for the respective research question.	**Yes**	Qual data is much more opinion and reflection. Quant data is about the actual market processes. So there is a valid relationship.
- Infrastructural resources considered	Feasibility of setting infrastructure to execute each method is considered.	**Yes**	Even if it is not all built, the feasibility is clearly assured.
- Compatibility with artefacts assured	It is established that the selected methods can help to progress towards the envisioned research artefacts.	**In progress**	The methods we currently execute well contribute to two potential artefact types: Shiny apps and knittr reports. Both are in our interest.
- Selection made	The list of relevant executable research methods is compiled.	**Yes**	

Figure 5.1: Filled-in scorecards for the "Research Methods" alpha with notes based on the commentary by Dr Bancroft.

> I think it clarifies it quite well. [...] Overall I think [that] the splitting of [the conditions] into the different components is quite logical and there's a clear rationale for it, and there's fairly clear progression through the different stages as well. [...] It's been a very useful process. I think it is particularly useful when we have multidisciplinary interests at work and when we've got outside stakeholders, who have to be coordinated with it. I think it's [...] very helpful for envisaging the whole project."

That being said, in the final interview of the case study Dr Bancroft provided an interesting critique of how that particular exercise was set:

> "I think if you have a team project, they should be using [the scorecards] in [a] team setting. [...] If you have a manager filling them in, it just seems rather didactic and certainly alienating for people. [...] If it's used in a team setting, it creates *team agreement*."

This last point indicates that, in Dr Bancroft's assessment, the value of the Social Data Science Scorecard Deck may transcend that of establishing "mere" shared understanding within a research team and provide the sense of *shared ownership* of the research process. This is important at the very least because the sense of ownership is linked to that of *responsibility* (Blau and Caspi, 2009) and thus may promote responsible research.

5.2.3 Identifying Issues for Revision

The use of the scorecards in the Dark Net project suggested a number of changes to the Scorecard Deck. While most of the changes were trivial (e.g. changes in wording for particular conditions), others significantly affected the contents of the Scorecard Deck and even some aspects of their use. The following discussion will describe and provide motivation for those changes.

5.2.3.1 Introducing prerequisites

The version of the Social Data Science Scorecard Deck used in the Dark Net project, as much as the original SEMAT Essence model (see Section 2.4.2), identified the state which each core project aspect (alpha) has achieved solely by checking whether the state's conditions are achieved. Therefore, the only way to tie together different alphas was through introducing 'interlinking' conditions. Such conditions were not my invention and are in the SEMAT model as well. For

example, the state "**Opportunity**: Identified" has a condition "The other **stakeholders** who share the opportunity have been identified" (OMG, 2015, p. 30).

Section 5.2.1.3 suggests that, when working through the scorecards of a single alpha, the interlinking conditions allow seeing this alpha as a part of the project as a whole and thus are valuable. However, in the project assessment session, when Dr Bancroft went through *multiple* alphas, the interlinking conditions gave raise to mild annoyance. For example, during that session he encountered the "Data" alpha whose first condition was "Research questions outlined", the "Research Methods" alpha whose first condition was "Broad research questions considered" and the "Research Questions" alpha that had an "Outlined" state – and all of those had already been discussed to some degree during the first session with the scorecards, as the "Research Goals" alpha also had a questions-related condition! Encountering the same condition multiple times leads to mechanical answering, while encountering a state that has been already considered as a condition in a different alpha (as indeed happened with the "Outlined" state of "Research Questions") may lead to frustration in regard to whether the earlier discussion has been detailed enough.

Dr Bancroft found this confusing. For example, when he encountered the "Infrastructural resources considered" condition in the second state of the "Research Methods" alpha, he seemed perplexed:

> "Infrastructural resources... *(pause)* So this [is] more asking about the practical feasibility of it, [...but] it's not under 'Infrastructure' though. [...] My immediate response is that it feels a little *redundant* [...], but I can see [why] you want to flag it. [...] My instinct would be not to have it [as a condition], but maybe just cross-reference that top-level text [from the 'Infrastructure' card deck]. [...Because when working through methods] people [would] think [...] more about the relationship between their data and their questions obviously. And I think that's kind of a different [...] thinking and a different mood."

In the interview after the case study, the issue of redundancy was the only serious criticism that Dr Bancroft expressed and hinted at what would become the solution to it:

> "I think there is a little bit of redundancy and I didn't get a sense of it linking things. [...] I wasn't sure why it was there. [...] A think that [...] a sense of project as a whole could come from cross-linking in other ways potentially. [...] [For example,] when you're off the "Research Questions", [...] another category like "Data" [gets updated] – like a traffic-light kind of system".

This traffic-light system takes the form of prerequisites, i.e. explicit requirements of achieving a certain state in a different alpha in order to progress the current alpha. The majority of states in the current version of the Scorecard Deck have prerequisites (see Section 4.3). In a printed version of scorecards, the benefit of prerequisites is in explicit pointing out that this is not merely a condition but a state of a different alpha. In a digital version (e.g. the interactive spreadsheet implementation), the benefit is in that the scores for prerequisites can be automatically updated and a user of the tool does not have to answer the same question twice.

One positive side-effect of turning redundant conditions into prerequisites was exposing the states in the Scorecard Deck that did not produce sufficient numbers of conditions independent from other alphas. For example, the old version of the "Data: Envisioned" state had only three conditions to start with, and one of them, as hinted above, was actually a prerequisite. This motivated me to engage with the fieldwork evidence again and to tease out the less obvious, but truly independent conditions for many states. This made the Scorecard Deck even stronger.

5.2.3.2 From "Research Methods" to "Analysis Methods"

When developing the "Research Methods" alpha in the early version of the Deck, I had a feeling that it was somehow weaker than the rest, but could not specify why exactly that was the case. The realisation came from Dr Bancroft's critique that he expressed after working through the first state of the alpha (see Figure 5.1). He was concerned with the use of terminology, as he expected to see conditions more related to designing the study as a whole and making sure that it was valid for the research questions, while instead he saw conditions regarding the choice of analytical techniques to which data can be subjected:

> "There's a slight difference between research design and method[s] and data collection. [...] There [are] preliminary questions about the research design that might be ones to add in. [...] So, you might want to have maybe a question about [...] internal validity or external validity and with reference to underlying object is studying. [...] That's perhaps distinct: you've got those questions, and then you've got [the] more practical questions about the methods, and then the process of data collection itself."

This suggestion, coupled with the existence of a validity-related condition in the next state, made me realise that the "Research Methods" alpha tried to be two things at once. It *did* put emphasis on specifically the analysis methods, but also incorporated some conditions in relation to the research design and validity – the ones Dr Bancroft referred to. However, most of the conditions

in relation to those overarching issues, within the Deck's model of research, naturally belonged to either the "Data" or "Research Questions" alpha. As a result, I made a decision to reduce the methods alpha to specifically the analysis methods and to give it a corresponding name, while making sure that the questions of research design and validity appear earlier on in regard to more appropriate alphas.

5.2.3.3 Allowing for more flexible progress tracking

Given my knowledge of the literature and the fieldwork, from the start of the Scorecard Deck's development I fully expected that this instrument will be used iteratively and with revisions of already progressed conditions and states as and when appropriate. For this reason, I went one step further than the original SEMAT model an intermediate level of condition achievement between "Yes" and "No" – "In progress". However, the Dark Net project showed that even *that* was not enough. As evident from Figure 5.1, our team exhibited a strong tendency to answer "Yes" to those conditions of which we had *sufficient* grasp. However, what exactly sufficed did change over time. As a result, in a brief progress tracking session closer to the end of the case study we revised many conditions that had already been ticked off, which did not seem quite right. However, the idea that we should have been ticking those conditions as "In progress" did not seem right either, since at the prior moment of marking them with a "Yes" we did not plan any immediate activities for progressing them and could move on to further states of the respective alpha. For this reason, the updated version of the Scorecard Deck distinguishes between the levels of completion "Yes (definitively)" and "Yes (to be revised)"[1].

5.2.4 Case Conclusions

The trial use of the Social Data Science Scorecard Deck was by and large a success, as it showed to be a useful tool for establishing team agreement both in terms of overarching goals and the particular actions to take – something of a particular difficulty in interdisciplinary projects. As Dr Bancroft points out:

> "[Scorecards] can be used as a [...] workable document across different disciplines, which is [...] something very hard to monitor and keep up with [...] because of that problem of project fragmentation that happens a lot just because people go back to their comfort zone."

[1] Additionally, an opportunity to dismiss a condition as Non/Applicable was added to current version of the Deck. This is not informed by any particular fieldwork observation, but rather by common sense and good practice

Contrary to my initial expectations, the Scorecard Deck appeared to provide the most value not in tracking project progress and thus making sure that all key aspects are getting progressed, but in establishing what those aspects would be in the first place. As Dr Bancroft characterises the later, progress-tracking sessions, those were "useful, but we kind of prefigured [a lot of that and simply] ticked the boxes." One possible explanation to that is the short span of the case study. If the project had the time to be developed more, there could potentially appear more interesting aspects to track. Another explanation is that we were a small team, so it was relatively easy for us to maintain close collaboration and common ground once the initial parameters of the project were set.

The spreadsheet implementation of the scorecards was very helpful, with arguably the biggest advantage being the ability to quickly add (and edit) notes regarding the discussed conditions. Besides, given the remote character of collaboration, using a physical version was not an option to start with. From my experience in social data science, remote collaboration in many areas is a rule rather than exception:

> "Having [the scorecards] physically located somewhere would be quite good for project management. [...] My thing is – I'm so *un*physically located in my work that I find them in a virtual work much-much better." (Dr Bancroft)

The case study also showed that the Scorecard Deck is reasonably intuitive in use and that the scorecard approach to social data science project management can be adapted quickly. That said, there is still a requirement for development of guidance that could help social data science teams to kick-start the use of scorecards.

> "People are obviously used to [...] narrative time-line. [...So showing how the individual alphas] are all linked together, but [...] not linear in the way [one] might expect – that would be very helpful. [...People would need guidance on] how you would return to things and update them." (Dr Bancroft)

Development of such guidance is part of the future work that follows up from this thesis (see Section 6.1).

Finally, the case study provided evidence that allowed to make several major improvements to the Scorecard Deck. The finalised version of the tool is assessed through broad evaluation in the following section.

5.3 Broad Evaluation: Expert Interviews

This section reports on the second stage of the evaluation – interviews with experts in social data science. The recruitment of the interview participants was done under two key considerations. First, the chosen evaluation mode put restrictions on the number of participants. The process of interviewing takes a lot of time, so a significant portion of those who expressed their interest in the Scorecard Deck as a tool to potentially consider in the future had to decline an invitation to participate due to schedule constraints. Besides, the population of social data scientists is still quite limited in size, as the discipline is relatively new and is arguably not as popular as some more computationally intensive and less interdisciplinary areas of data science (for example, computer vision and speech recognition as indicated in Section 2.1.1). Second, the qualitative nature of the evaluation asked for "qualitative representativeness" (Diefenbach, 2008, p. 879) of the findings – i.e. focusing on representing the qualitative variety of perspectives rather than on mimicking the distribution of perspectives in the whole population of social data science experts. Given the rapid developments in social data science, qualitative representativeness is more important than the statistical one, as what is peripheral to the field today may become mainstream tomorrow.

Most of the potential respondents were identified and approached systematically through relevant research associations (the Oxford Internet Institute and the Alan Turing Institute) and publication targets (the Big Data & Society journal and the SAGE Handbook of Social Media Research Methods – cf. Voss et al., 2016). Some potential respondents were contacted based on their wider knowledge of the field, which allowed to invite social data scientists who are employed outside academia (e.g. representatives of CASM – "dedicated digital research hub"[2] of the Demos think-tank). Finally, one particular respondent was a member of the a bank's data science team, as their colleague had expressed interest in a demo version of the scorecards presented at a SICSA DemoFest (SICSA, 2017).

Every potential respondent, including those identified as part of systematically approached groups, were contacted through personal emails rather than trough bulk invitations. While this approach limits the number of people who can realistically be contacted, it is also known to increase the response rate (Sahlqvist et al., 2011) and has a significant advantage in control over respondent profiles. While the nature of the considered academic groups already assured some related expertise, additional respondent selection had to be made. For example, some of the identified academics did not actually do data science, but studied the process (e.g. through non-participatory fieldwork) or its implications (e.g. ethical or political). Others did data science, but not *social* data science. To ensure that only the respondents with first-hand experience in

[2] https://www.demos.co.uk/research-area/casm/. Accessed: 2018-12-11.

doing social data science were interviewed, I screened person's publications, research projects and their stated research interests (provided that the latter are described in sufficient detail).

The level of participant seniority was another concern. Indeed, the proposed research exercise is an expert evaluation, and the notion of "expertise", in its lay sense, may be necessarily associated with long-standing presence in the field and some indication of high position within it. Yet, expertise is a challenging concept (Fischer, 1990; Ericsson and Smith, 1991; Jasanoff, 2003). Collins (2013) demonstrate the dimensions in expertise and argues that its key aspect is immersion in the community of other experts. In the case of social data science, as a relatively new discipline, the "junior" community members may be no less immersed in it than the "senior" ones. Moreover, less senior team members may have more hands-on experience in doing the research work, while the more senior team members might focus on management. Finally, the anecdotal evidence of the case studies suggest that it is not unusual to have social scientists as senior members to define the project direction and junior members with technical backgrounds to implement the required analyses. This can give varying perspectives, both valid for the evaluation needs. That being said, the experts should have experience sufficient to critically reflect on their work, to argue about the causes of its successes and challenges. For this reason, all the potential interviewees in academia were at least PhD students in the last year of their study or had demonstrable prior industrial experience.

5.3.1 Relevance of the Scorecards

The respondents, who across all of them have worked through- (or been asked open questions motivated by) the contents of the scorecards in relation to every alpha in the Essence of Social Data Science model (see Section 4.1), have uniformly agreed that the points suggested by the scorecards can be relevant in a social data science project. In most cases, when reflecting on a suggested point, a respondent came up with a relevant example from a project they had chosen as a scenario. For example, John[3], who worked through the "Data Sources" alpha using his prior project in data-driven epidemiology as an example, offered a lot in regard to the conditions suggested in the first state of the alpha – "Data Soures: Indentified" – reflecting on the suggestions to consider demographics, triangulation and use of traditional methods (see Table 4.6):

> "We wanted to gather the age of people as one of the parameters, wealth the people one of the parameters, [and the] geographic location is the fundamental thing. We got interviews. We actually had a parallel project that we didn't conduct but there were a psychology group who did interviews with people in [a different university].

[3]Hereinafter, all the names are pseudonyms.

> [They were] studying the sort of same demographic [groups] but doing it with classical social science to see what people do to avoid catching flu [and to avoid] places [where, as] they heard, other people had it. [...] So [what] we were doing is directly measuring that via proxy [of] location [and] they were doing it by going around interviewing individuals and groups."

When switching to the next state – "Data Sources: Evaluated" – whose conditions and prerequisites suggest different points of critique for the identified sources (see Table 4.7), John could offer just as much and was appreciative of the fact that compliance-related issues were defined as a prerequisite:

> "Level of control, access, accuracy of data are all fundamental. [...We] really needed to know about the veracity of the data. How to assure that was not obvious because of quite an early experiment in using smartphones for the for epidemic tracking. [...] And then compliance – [the senior] medical ethics people obviously required [that from] us, you know, because this was part of a medical epidemiological study we had to we had to have very good privacy."

As much as John acknowledges the value of complementing new data sources with the traditional ones, Joseph does the same for analysis methods (as suggested in Table 4.17). In particular, he explains how qualitative analysis of interview data helps him to design quantitative analysis of the mobile device logs (i.e. times, lengths and directions of mobile phone calls and text messages between participating individuals):

> "When you actually talk to people, they say things that perhaps you haven't expected and so what I do you in a more qualitative approach to that data is read through all the interviews, read through the coding of the interviews [to] get a sense of, like, well, what types of relationships would [...there be]: something about the type the nature of relationship, [...] about the context in which this relationship exists in terms of where people are located – physically at home, or work and so on. I go through all that process. I would sort of take out different types of relationships depending on that general overarching interest [...]. What I want to do [...] then is sort of like a, some of the quantitative aspects of the calls and texts [...]. I would go to the log data and I would compare that to the other types of ties more broadly within the particular respondents' call and text log. And then I would also potentially, if I thought it would be useful, compare those patterns in the much larger national dataset."

Some points suggested by the Scorecard Deck provoked extensive discussions with different evaluation participants, thus further affirming the relevance of those points. For example, the condition of envisioning proxies and metrics resonated with Jane and Nina who both chose to work through the "Research Questions" alpha (see Table 4.13). Interestingly, while both Jane and Nina talked about projects that had employed machine learning for a predictive analysis problem, the challenge of metric choice had been most pressing for them in different circumstances – Nina had to choose a salient proxy to represent an obscure variable, while Jane had to choose a metric to evaluate the quality of her models:

> "The second part [of the discussed project was] analyzing the image content and then basically try[ing] to see whether we can actually match it with – well, I would say the number of protests that [are] happening around the world, but there is no database, so we use[d] proxies for ground truth, which is the Guardian articles. [...So,] proxies and metrics envisioned – yeah [, that was important]." (Nina)

> "In the dataset [...], we have one majority class [...]. Therefore, if you're training [a] model [and] it strikes for high accuracy, then the model that always chooses majority class is doing pretty well [...]. Therefore, like, the choice of metrics was quite important to this project and we indeed actually compared several choices." (Jane)

Some respondents also commented on the relevance of the suggested states as a whole rather than only on that of individual conditions. For example, Phillip recognised the "Team: Seeded" state (see Table A.20), as in his project there was "an initial set of people who were the kind of backbone of the team", however those people "also knew that [the] initial team didn't have the necessary skills to do everything that [the] project would involve". Similarly, Vassiliy, while discussing the project of adapting the tools for graph analysis available on the market to prevent credit card fraud, recognised the stage of formulating the research questions as a broad framework (see Table 4.12):

> "We do a foundations work [...] when we would look at [...] a number of tools. We wanted to look at Neo4j[4]. [...] We ended up looking at a Python Graph Model and we did research around that subject."

In many cases, the points suggested by the scorecards seemed to make the respondents conceptualise some of the process that were happening tacitly during the work on their project as

[4] https://neo4j.com/

a by-product of the collaboration rather than as a result of a focused effort. For example, Phillip explained how in his project the selection of a management model (suggested in Table A.20) had never been explicitly done, yet effectively a flat hierarchy model (cf. Carzo Jr and Yanouzas, 1969) had been employed:

> "We never thought of it, I guess, as a management model, [...] that was perhaps putting it a little strongly for us. But we had this idea of there are the leaders, [...] there are the PIs and [the] Co-Is of the research, and the research associates. But the research associates are effectively proposing projects to the PIs and Co-Is and then in [...] face-to-face meetings we all kind discuss what the value of these projects were, we would refine them together. So, there was a leadership in that we took guidance from the PIs and the Co-Is, but it was a little bit flatter than [otherwise possible]."

These suggestions in the scorecards that aid in conceptualisation of project aspects may be of value even if they do not motivate the users to achieve something completely new. Indeed, simply by making something explicitly, it may be easier to manage this aspect and also to pass on the derived practices to new projects and/or team members.

In some cases, the respondents acknowledged relevance of the conditions suggested by the scorecards with a caveat that in practice achieving such conditions is difficult or even impossible. For example, Juan, when discussing the "Analysis Methods" alpha and commenting on the suggested condition of knowing how to evaluate the confidence in the results produced by each of the selected methods (see Table 4.17), had to assume a very cautious position. In his judgement, to truly estimate the confidence in the findings derived with modern analysis methods and from new data sources, a wider culture of- (and a corresponding compliance framework for-) reproducibility in social data science is required:

> "I'm not so sure about that. I think there's an assumption that if someone else has already used those methods and [they] have been socially validated – and by 'social' I mean the results and the articles have been published in peer-reviewed journals – you have confidence. But I don't think that we have enough in history – or even any experiences in digital humanities our data-[driven] social sciences – [...] related, for instance, to the replicability of results. So, I don't know of many studies in, which people have gone back use the same data set and come up with a different conclusion. We don't do that yet. [...] I mean, in many cases still there are not so

> many datasets published. [For example, the often used] Twitter data, in most cases, precludes you in its terms and condition from republishing the dataset."

One particular point that most participants acknowledged as relevant only in a hypothetical case rather in the project they used as a scenario was stakeholder engagement. This point appeared in discussion with many respondents, as stakeholder engagement in some shape or form appears as a prerequisite across states of different alphas in the Scorecard Deck, e.g. for outlining the research questions (see Table 4.12) and for envisioning and iterating the artefacts (see Tables 4.22 and 4.24). A common reaction to this among the respondent was to think of their project funders and claim relative independence. For example, see the following two quotes from different participants[5]:

> "In my case with this particular grant the language [of] the funding is that as long as the project is in the general spirit of the proposal that's fine. [...] I'd be quite honest and frank with you [...], but this is how research, I think, goes for a lot of people."

> "We were told [by the funders that] this [project] was about trust online and kind of how trusted feed into conversations [across] various online domains, but [...] we had full control over what those domains were. So one of them was to do with welfare and poverty and social media, one of those to do with kind of more traditional media forms. [...]. So, we had a lot of choice and even though, this was a funded project by a combination of funding bodies, [...] but the stakeholders didn't really have any kind of say in what we chose to look at."

Sometimes I directly asked the respondents about stakeholder groups beyond funders that may be less obvious in the context of the term, but included in Freeman's (1984) definition. For example, I asked Nina whether she, when planning her research, considered how to make it more valuable for other academic stakeholders – e.g. social scientists who also study protests. She said that this was something she would like to do in the future and that she had "proposed in [her] thesis that one potential expansion would be including a social scientist". However, even in this case, as she rightfully acknowledged, that would actually be a "collaborator [...] on board".

Such an attitude can be partially explained by the fact that many respondents only saw the stakeholder-related prerequisites in other alphas but did not interact with the "Stakeholders" alpha itself. The only respondent who *did work through* this alpha, John, commented: "That's definitely [relevant], that list [of conditions] like tacit stakeholders accounted, [...] getting people

[5]These quotes are specifically sensitive and thus are not attributed to specific respondents.

involved". Partially this may be caused by the nature of John's research project – epidemiology is a sensitive field and at the very least the interests of the patients have to be accounted for. However, arguably another reason is that the "Stakeholders" alpha is another one that helps to conceptualise otherwise hidden concerns. Therefore, only through engaging with the unexpected points suggested by this alpha (e.g. inclusion of tacit stakeholders) it is possible to appreciate its importance.

Furthermore, even if the "Stakeholders" alpha is not applicable to *every* social data science project, I would argue that its inclusion is still valuable as sometimes it is *very* important. For example, Chris talked about how for four consecutive years, he had been recruited as part of a big social data science team for several days every summer to provide real-time data-driven analytics for a major annual music festival in Europe with the aim of improving customer experience. For him, stakeholder-related issues were of major importance, as their requirements would often fluctuate:

> "I often have this picture [about this project] that you walk into a forest and you have no idea what you're going to find. [...] I think it's kind of natural that it is [a bit chaotic], but [it is] also the unpredictable nature of the stakeholders. [...] Depending on what's happening on the journey inside the forest, [...] it's quite unpredictable to say [...] what they want to dig into in one or two or three days."

All in all, the evaluation shows that the content of scorecards is relevant in different kinds of social data science projects. The scorecards do indeed point at explicit and tacit project aspects that are worth managing and suggest conditions that are beneficial to achieve – even though some of those conditions are applicable only in subset of projects (e.g. those where the role of external stakeholders is high) and others are almost impractical to achieve at this day and age, but arguably may become possible in the future (e.g. thorough evaluation of confidence in outputs of different social data science research methods).

5.3.2 Clarity of the Scorecards

By and large, the respondents did not face any significant issues in understanding the contents of the scorecards. The conditions proved to mostly be formulated sufficiently clearly for the respondents to swiftly get their gist, yet sufficiently flexible to be successfully interpreted in the context of different social data science projects. For example, within a standard hourly interviews, Jimmy easily worked through all states of the "Data Sources" alpha. John only worked through the earlier states of his alphas, but he managed to consider three of them – "Stakeholders",

"Data Sources" and "Compliance". Juan, who only had 15 minutes to actually work through the scorecards, managed to fully cover the first state of the "Analysis Methods" alpha.

Occasionally, the respondents experienced difficulties with making sense of the alphas that had been out of scope of their involvement in the discussed projects. For example, this was the case with Jane who chose to talk about the "Research Goals" alpha as a logical place to start rather than because it was the most familiar project aspect for her ("Oh, well, should we start with research goals – like, from the beginning?"). Her project was focused on using a particular research technique (conversation modelling) to achieve one well-defined objective – to improve classification of stance towards rumours (belief versus doubt) in Twitter discussions. In her work she did not need to put much consideration to what the real-world problem that motivated her work was, what the interests of the stakeholders were and to other similarly higher level issues. This was not because she did not believe those were important – in fact, she eventually talked about those and contextualised her work accordingly:

> "if we [...] kind of zoom out and abstract a bit, [...] the problem is rumour spread online. [...] And we want to tackle it, so we ideally want mathematical methods that will assist in rumour verification. But we started to looking at stance [...] This was motivated by previous research that has shown that [...] rumours that attract a lot of scepticism [...] are more likely to be proven false."

However, according to Jane, the wider problematics of her research were relevant on the level *beyond* her project and rather on the level of a whole strand of work of her research group. For her, the research goals were given rather than worked through.

Several respondents experienced slight confusion when trying to distinguish between consecutive states within an alpha. For example, when looking at the "Research Questions" alpha, Vassiliy found the first two states – "Identified" and "Refined" – to not be completely discrete. According to Vassiliy, "in a research project [questions rather] evolve". This, however, is not necessarily a pitfall of specifically the Scorecard Deck. Indeed, any project management tool that tracks progress offers abstractions for stages (be that states of project core aspects, steps in a process pipeline, etc.). These abstractions have to be discrete and thus cannot fully reflect the continuity of reality – hence, some overlap between the consecutive ones would be inevitable. It is thus more important to make sure that separating two consecutive states actually adds value. In Vassiliy's mind, that was the case. Indeed, he could distinguish between the two overlapping states by pointing out how the initial broad question of his project (whether "graph[s can be] helpful in understanding common point[s]-of-compromise") was refined to the level of the specific graph

analysis methodologies that his team had looked into. Moreover, after reading the individual conditions suggested in the state-level scorecards, Vassiliy agreed that "all of those would be included" into consideration within the project work.

Yet, some other participants felt a bit stronger about the overlap between consecutive states, especially when their respective conditions also represented progress in overlapping issues. For example, Jane felt that the "Artefacts" alpha has, across its states, too many related conditions:

> "I think I am overwhelmed by the amount of questions [presented as conditions in the scorecards] and that [...] sometimes, when you answer, you feel like you have already answered it for certain questions. [...] [For example], I guess identifying artefacts in advance is [...] a good thing. [But], I mean, I'm not sure you need, you know, like, a thousands [questions on] identifying artefacts. You just say to yourself: well, I wanna publish a paper, [...] and maybe these are my venues I'm going to publishing in, and these are my deadlines."

To an extent, this feeling of being overwhelmed by overlapping conditions is related to the design of the evaluation. In reality, the states of the scorecards are meant to be filled in alongside the managed social data science project actually getting progressed, whilst in the evaluation interview a respondent has to fill the cards in retrospectively. Thus, it may be quite natural for a respondent, even when they see the conditions related to the earlier states of an alpha's progress, think of the later states. Under this scenario, when this respondent arrives to the cards of the later states, the new conditions may seem redundant. Actually, Jane also felt this way, and said that she "might be interested in actually in [...] reading the questions and see[ing] if that helps" in an "ongoing project" that she had at the moment.

However, as suggested by Nina, the confusion between the overlapping states may come from the different ways in which the Scorecard Deck suggest to represent the progress of an alpha and the progress of its particular conditions. Indeed, progressing an alpha is intended to be represented by going through the alpha's states and filling its conditions one-by-one, while progressing a certain condition is meant to be represented by selecting different states of achievement for them ("Not achieved", "In progress", etc.). According to Nina, a user may get confused when to select each one:

> "I think [...] the ['Research Questions:'] 'Outlined' and 'Refined' part had a lot of overlap. [...] It kind of [...] makes sense, but the reason I think the overlap is a bit confusing is because you have this [...] 'In progress' [as an option for how strongly

> a condition is achieved]. So, yes, you kind of come up with an outline and then you refine it and then your [...] 'In progress' becomes 'Yes (definitely)'. And too much overlap basically kind of overkills that 'Yes (to be revised)' or 'In progress' in the 'Outlined' part."

I plan to address this critique in further work. A simple brute-force approach to it would be to cut on the number of states and their respective conditions. However, such an approach is sub-optimal. There *is* a value in overlapping conditions since in more complex social data science project the subtle differences between similar conditions may appear to be of critical importance. Moreover, they allow for a more flexible and rigorous progress tracking. For example, Phillip was very appreciative of this continuity in conditions and how elaborated they are:

> "I think if I was kind of using [the Scorecard Deck] at the start of [our project] – so, like, four years ago – [...] we'd be able to see the [conditions] that we couldn't [even start to] get addressed. And that would be useful for us as well. So the fact that we wouldn't be able to answer some of these questions [would] kind of [give] us ideas on 'Where should we go? what should we do? By the next time we have meeting in six months, let's try and fill in those gaps'. And so that in itself will help give shape [to the project]."

What, however, can be suggested to remedy the critique of redundancy without corrupting the value of the the Scorecard Deck, is to allow further customisation of the deck for the needs of a particular project. In the current implementation of the scorecards as an interactive Google Sheet, a user can only customise the list of alphas they would like to consider in their project. This can be further extended by allowing users to either completely de-select particular states of certain alphas from consideration – or to blanket-select a certain state to be completely achieved without looking in detail at its individual conditions. The latter would be especially useful for such project as the Jane's one – when there is enough set-up information at the very start of the project. In such a project, for example, broad research questions may already be defined, so the project members can jump straight into their refinement.

Additionally, development of a lighter version of the Scorecard Deck can be proposed. This version would take a form of simple checklist of the most pertinent and qualitatively distinct conditions from the Scorecard Deck. Under such approach, the conditions could still be grouped by alpha, but the individual states of each alpha would arguably collapse. Such lighter version of the Scorecard Deck would thus not be suitable for tracking progress of a social data science project, but can still be valuable as starting discussion point when planning a social data science

project. It also can be a good introduction to the lingo and contents of the scorecards and thus may be advised to use alongside the full deck when first using it.

Finally, in very rare instances the respondents seemed to misinterpret the very gist of some of the suggested conditions. The only explicit example of this comes from Nina who did not interpret some of the earliest conditions in the "Research Questions" alpha as intended. For example, she interpreted the "Demand Understood" condition (see Table 4.12) as that a project team has all the sufficient knowledge to conduct research, while the intended meaning was much weaker – that there is an understanding what kind of new knowledge the research project *aims to produce*. While I would like to re-iterate that such misinterpretation was a rare case displayed by just one participant, it is still something to consider in further work. One specific action that can be suggested is to revise the language of those points in the scorecards that have caused confusion. More systematically, a glossary for the terms used in the scorecards that would provide more detailed definition for each condition, possibly with some examples, would be of great value.

Overall, the Social Data Science Scorecard Deck proved to be an easy-to-comprehend project management tool. It only caused confusion with some selected respondents, it was never confusing to the level that actually made it unusable or pointless, and some of this confusion would most likely disappear in a real-world scenario of an ongoing project, especially if the tool was used continuously and thus the team could go up the learning curve. That being said, the fact that some confusion is still there suggests that further guidance for the use of the deck – e.g. a user manual – should be compiled. In fact, this idea is supported by Juan, who assessed the comprehensibility of the Scorecard Deck as follows:

> "I think that for someone who has used before [any] project management tool, [the scorecards] would be very intuitive. For people who have not – they will need some clear instructions."

5.3.3 Comprehensiveness of the Scorecards

So far the discussed respondent feedback had been focused on whether the contents of the scorecards clearly relate to their own work and to social data science in general. The question that is still open though is how complete the Scorecard Deck is – i.e. whether making sure to fulfil the conditions suggested by the Scorecard Deck is enough to be confident in successful completion of a social data science project. A relevant analogy would be that to precision and recall (cf. Buckland and Gey, 1994) – while the former is already assessed, the latter requires further discussion.

Unfortunately, reliably assessing the scorecards' recall without actually applying the tool in a large portfolio of research projects is problematic. Indeed, it requires thinking of the types of issues that have not been identified neither during the fieldwork nor in the literature, which may be difficult without actually facing these issues in practice. The expert evaluation allows to partially achieve the same effect, as the respondent may bring up the unconsidered issues from their own experience. However, if a respondent is asked to simply work through a particular set of scorecards (which was the case with most evaluation interviews), it creates a risk that this respondent would be distracted from the issues not covered in the scorecards. To compensate for that, I conducted the interview with Joseph differently. While I still asked him some seed questions inspired by the alphas of the Deck, no focus on the Deck's particular conditions was put. I encouraged Joseph to drive the conversation with the help of appropriate probe questions (discussed in Section 5.1).

By far and large, the accounts of doing social data science provided by Joseph highlight the issues that are captured by the Scorecard Deck. In fact, Joseph's project presents an excellent example of how the issues captured in the conditions of the "Data", "Data Sources" and "Compliance' alphas (see Sections 4.3.1, 4.3.2 and 4.3.6) postulate themselves in a social data science project, and how these project aspects affect each other and the project infrastructure (see Appendix A.3) in accordance with the prerequisites of their alpha states.

As Joseph's project was concerned with discovering patterns in human mobile communications (i.e. calling and text messaging), he required data on numbers of calls/texts between different participants, their frequency and directionality. To contextualise those, he also required data on the type of relationship between communicating sides (friends, classmates/colleagues, relatives etc.) and on the participant demographics:

> "[I wanted to look at] how adolescents use mobile phones and whether or not their use of mobile phones tends to result in [...] forming these insular small group relationships – one or two close friends [who spend] all their time texting and no time doing anything enriching for their lives. [...] Or if, in fact, they're texting a lot but they [are] having a series of intense relationship, [...which is] quite different than what adults experience."

This presented a challenge of accessing publicly unavailable data and satisfying the participants' privacy concerns. This was done by developing an Android app:

> "[We developed] an Android application that allows us to collect non-identifying calling and texting log data – [...] just hash identifiers associated with phone numbers,

> time and dates of calls and texts – but also to be a research tool. [...] Respondent, [when] installing it, [...] knew exactly what it was doing and it could be woven into onscreen surveys where during the survey [the app would], say, choose the most texted person from their address book and ask them a question about [the recent call] or about that person."

Joseph and made sure that "no contacts, no text messages – nothing" of specifically high sensibility were accessible to the research team – for example, numbers were hashed on the app side "so that they would never leave the device". As Joseph characterised the process:

> "It was almost overly precautious, but I think it was important."

As can be seen, in accordance with the conditions of the "Data: Enivisioned" state (see Table 4.1), Joseph required evidence on primary (e.g. number of texts between participants) and secondary characteristics (e.g. the demographics of the texters) of the studied phenomena. As shown in the "Data Sources: Identified" state (see Table 4.6), this could affect initial selection of data sources – and indeed, the qualitative difference in data on primary and secondary characteristics motivated motivated him to consider traditional data sources (surveys) alongside modern ones (call logs) and the private nature of the data motivated him to think of sources with leveraged access – all in accordance with the state's conditions. In his thinking about data acquisition, Joseph had to account for compliance considerations in accordance with the next state of the "Data Sources" alpha (see Table 4.7). The resulting solution came in form of adding an extra element to the project infrastructure architecture – a mobile app – the development of which was tailored to the considerations around the data sources as suggested in Table A.14.

One of the leitmotivs in Joseph's narrative is how his work is not simply iterative – it is in a sense never-ending, with new projects stemming out of the old ones:

> "[After my current strand of work] I have a book proposal that I'm putting together, which will really tie all of that data [I am studying currently] together. [...And] I think I'm going to be harvesting, so to speak, this [type] of data for years to come. I think can get number a of other [...] papers out of this data that are significantly different than what I put in a book. So, it's sort of like, in a sense, [this work] isn't over."

While the scorecards do pick up on this permanent continuity, as the latter states of some of the alphas explicitly talk about planning further research (for example, see Tables 4.20 and 4.16)

and about assuring the possibility to reuse of the data, tools and ways of working (see Tables 4.5, 4.11, A.19 and A.35), there is a specific detail that emerges from Joseph's narrative that is not explicitly covered in the Scorecard Deck. When Joseph talked about what specifically could *spark* the new work, he stressed the importance of communication with potential future collaborators:

> "[Me and a colleague] [a]re talking at a conference and [...] he mentioned that he had, you know, some survey data that he'd be interested in [...] working on with me. [...] And I, at that time, had been thinking a lot about [mobile phone] log data, and digital trace data and sort of advantages and disadvantages as compared to more typical types of measures of this particular media use within social science. [...] I [was] just talking at a conference, that [idea] just came about – so it wasn't really premeditated in any real way."

Even though the process of arriving to a research idea discussed in the quote above is spontaneous, the key steps that lead to it – networking, discussing ideas with other professionals in the field – can be taken systematically and thus can potentially be incorporated into a project management tool. Specifically to the Scorecard Deck, related conditions can be added to the early stages of the "Research Goals" (e.g. "Input from other researchers in the field received") and "Team" (e.g. "Networking with other professionals in the project's field performed") alphas.

It is worth noting that unforeseen issues that are caused by unique circumstances of a particular social data science project may still arise. For example, one the interviewees[6] discussed how his project team decided to collaborate with a certain public figure who was a convicted and how this collaboration backfired in an unpredictable way. For a reader, it is arguably clear that a case like this is specific to the project and not characteristic of social data science in general. The Social Data Science Scorecard Deck does not only fail to deal with such issues – it cannot be expected to do so. Therefore, regardless of how complete it is, it still cannot replace the informed judgement of a project team on how complete their project is.

5.3.4 Helpfulness of the Scorecards

Most of the researchers suggested that the scorecards are likely to be especially useful when the team is interdisciplinary and large, because there are more relationships to manage (which make a project management tool more desirable to start with) and because the Scorecards can serve as a point of common lingo. Even Joseph, despite his remarks on potential detrimentalism of using

[6] Due to sensitivity of the issue, neither attribution to a particular respondent nor direct quotes are provided.

project management tools in research work, thought that the Scorecard Deck can be a valuable talking point:

> "I have worked with other people from different disciplines and there's often a disconnect in terms of language. [...] So having something that's tangible that you could both discuss... – I mean, you still want to make sure that [...] the team members really discuss each point thoroughly. I think that would be very critical to [the Deck] being successful. If team members are just like, 'Okay, I got this covered, you do that one over there', then you'd have these misunderstandings quite a lot. But if the team members all together sat down and went through each one [scorecard] at a time, then I think it could help provide sort of common language and avoid misunderstanding. So, in that situation, I can see [the usefulness]."

Nina shared the same stance and suggested that when a social data science team is relatively small and tech-focused then, a generic project management tool such as the Scorecard Deck may be an overkill for tracking the project's progress. According to her, if "you are writing code together with the entire group, then probably I would switch to GitHub [7] because it would be much easier to track [progress...] with GitHub's project timelines", while in a more mixed background team, some of the members "won't understand GitHub. They won't find it user-friendly" and thus the scorecards would be handy.

It is worth noting that an instrument where a project team has to define its timelines, milestones and deliverables (including that provided by GitHub) is only useful to manage an already operationalised project, but not to design a project. The Scorecard Deck's ability to *provide guidance* to a project team cannot be replaced with such tools. In some circumstances, this may not be required: the limited evidence provided by Vassiliy suggests that in an industrial setting data science projects are often quick and straightforward, e.g. refitting old models with new data. Moreover, as Vassiliy says, he does not believe that Tesco Bank does "out-and-out research projects", as most of the projects have "clear outcomes" and the primary concern of pretty much all of them is "looking for financial benefit". Yet, even the Tesco Bank data science team frequently faces more open-ended challenges in the form of "proof of concept" projects where "the financial expectations or the outcomes are less clear". Such project most often study the *efficiency and effectiveness of the employed data science methods*. It is the open-ended projects where Vassiliy suggests that the Scorecard Deck is especially useful.

[7] https://github.com/

Jimmy takes a balanced approach by suggesting that the Scorecard Deck can be valuable in both individual and team work, but in the team setting it can add value in more ways:

> "I can see it being very useful for individuals because I would bet that no single [researcher] put[s] [all relevant point] in the list. [...] So, if I were in the beginning of the project [that had been discussed during the interview], honestly, [the scorecards] would already kind of help me get a *very*[8] detailed understanding of kind of things to cover. [...] But for teams I find that [the deck] can be a very useful learning tool. So, if I showed this to my political science friends, they would probably look at them and think, 'Oh, I'd never think of that!' Just like I would go through and [find a] side [that] didn't cross my mind."

This last property is a good response to John's critique that only part of the scorecards' content is specific to social data science – a lot of issues are common for social science in general. This is hardly a problem if the scorecards are able to teach social data scientists with technical background on the common problems that come from the social science side and vice versa. In face, because of this last property, Jimmy suggests another context for using the Scorecard Deck – *education*:

> "[For the] future students of mine, [the deck] would be useful, [as they would] understand what a social data science projects involved and particularly how [they are] different from regular data science. [Regarding the] Masters course in Social Data Science in our department [...], a lot of [...] computer scientists [in] the programme [have] never worked with human data or with any kind of, like, more solid data. And the people who come from the social sciences, they probably won't think of things like monitoring and kind of API access, [...] restrictions and even the computational needs of a project. [...] I'm gonna be supervising some students next year. I think I'll show [the Scorecard Deck] to them."

[8] Emphasis based on intonation.

Chapter Six

Conclusions

This thesis meets the objective set out in the Introduction and adapts a project management tool from software engineering, the SEMAT Essence model (see Section 2.4.2), to social data science.

In regard to objective 1, I establish several dimensions of challenges that social data science faces. First, the technological innovations that underpin social data science originate in a close circle of big hi-tech companies. Even though some means for doing social data science are getting increasingly available for a larger number of actors, the drastic inequality in access to data about people and their behaviour and to technological means of their analysis still persists. This makes social data science an instrument of political power and raises significant issues of ethics and responsibility in social data science. These issues are amplified by the epistemological challenges that arise from the mismatch of contemporary data analysis methodologies with the research questions typical to social studies, as well as from the nature of the data sources.

In regard to objective 2, I make a choice in favour of adapting the SEMAT Essence model. This tool does not assume or suggest any particular work process and rather focuses on objectives. This is desirable given the variety of social data science projects. Moreover, the SEMAT model provides holistic guidance that covers both "hard" (technical) and "soft" (managerial) issues and addresses concerns common for social data science such as stakeholder engagement and team management.

In regard to objective 3, I conduct four case studies in social data science. Each case presents a single social data science project. Together, they cover a broad spectrum of areas within the discipline: impact-oriented academic work (cf. the Dark Net project, Sections 3.4.2 and 5.2), social media research consultancies (cf. the Shakespeare Lives evaluation project and the InfoMigrants evaluation projects, Sections 3.2 and 3.4.1) and production of a data-driven radio show (see Sections 3.3). The case studies were done at different points in time within my studies.

They serve different purposes and are treated differently in the write-up. The two case studies presented in Sections 3.2 and 3.3 are complete ethnographies. They provide a holistic picture of what it takes to be doing social data science and provide account of all significant issues that I observed and, as a team member, experienced. While the write-ups of the other two case studies are not as fully-fledged, they report on the observations that provide significant added value. Overall, the four case studies present **a contribution to understanding of the challenges involved in doing social data science**.

In regard to objective 4, I develop and present the Social Data Science Scorecard Deck, a holistic project management tool for social data science that borrows the basic structure and content from the SEMAT Essence model, but reworks it in three significant ways. First, the Scorecard Deck adds six sets of scorecards for the core aspects of social data science projects that are systematically informed by the fieldwork and the literature. The other six sets that are based on the SEMAT model are appropriately modified. Second, the Scorecard Deck adds a new structural element that is absent in the SEMAT model: the prerequisites. They represent how different aspects of a social data science project are related and how progression in one project aspect may rely on progression in some other aspect. As such, they truly support holistic project management. Third, the Scorecard Deck abandons the paper representation of the scorecards in favour of a digital one as a Google spreadsheet. Through argument and, subsequently, in-depth evaluation I show that digitisation of the scorecards is more than a mere change in their presentation – it actually affects the way the scorecards can be used and makes them a significantly more powerful tool specifically for managing social data science projects. As such, **the developed project management tool contributes to good research practice in the field of social data science**.

In regard to objective 5, I take part in a social data science project whose investigators agreed to use the Social Data Science Scorecard Deck for project management. I foster the use of the Scorecard Deck in different settings and for different purposes. Throughout the resulting series of participant-observation exercises, I establish the value of the Scorecard Deck for shaping project goals, for establishing and maintaining shared understanding and team agreement and for making sure that the team members with different disciplinary backgrounds work towards a common goal. I also establish several issues with the Scorecard Deck. Some of these issues – for example, the tension between linking scorecards and their partial redundancy – only become evident when results of multiple exercises are considered together. Thus, I **contribute the methodology that can be used for further fine-tuning of the Social Data Science Scorecards Deck**.

In regard to objective 6, I subject the Social Data Science Scorecard Deck to an external validation through interviews with experts in social data science. By asking the interviewees to retrospectively apply the Scorecard Deck to one of their past projects, I make sure that the

applicability of the scorecards is assessed in respect to qualitatively different research projects. The assessment confirms the value of the scorecards, however suggests that modes for easier customisation of the tool are required, as the comprehensiveness with which the scorecards go through often overlapping points may appear excessive for many studies.

6.1 Limitations and Future Work

While I believe that my work is substantive in volume and robust in methodology by standards and reasonable expectations that can be applied to a PhD thesis, it is necessarily limited by the number of case studies – especially since only one of them was used for in-depth evaluation – and the number of interviewees who participated in the broad evaluation. Furthermore, I think that the Scorecard Deck in its current version may overfit to social data science as done by academics. Again, this is in the nature of a PhD and the opportunities for doing non-academic case studies. Even though the informing case studies were not academic projects in terms of their research goals and produced artefacts, they all included academics on their respective teams. Out of the interviewed experts, only one is employed outside academia.

Consequently, further research will need to address these limitations by further validating the tool in other contexts. This includes exploring the possibility of adoption of the tool in contexts such as industrial work or, for example, in civic hacking or data journalism. Further case studies or other forms of validation will no doubt generate more changes to the tool and its contents.

Another aspect that could not be developed as part of this PhD was to interact with the SEMAT team. As mentioned in Section 2.4.2, the original SEMAT model is published as an OMG standard (cf. OMG, 2015). In the future, it may be valuable to contact the SEMAT team to explore whether the Scorecard Deck can be incorporated into the standard as an extension.

Finally, further dissemination of the Scorecard Deck would require development of accompanying usage guidelines and other educational materials in order to make the tool fully usable without external supervision.

APPENDIX A

SCORECARDS ADAPTED FROM THE SEMAT KERNEL MODEL

The scorecards for each alpha in this group derive their content from some alpha in the SEMAT model – in most cases, it is that alpha whose role in the SEMAT Essence model corresponds to the role of the informed alpha in the Essence of the Social Data Science model. For example, as discussed in Section 4.1.1, the "Research Goals" alpha in the Essence of Social Data Science corresponds to the "Opportunity" alpha in the Essence of Software Engineering – and thus the scorecards for the former are informed by the contents of the latter. An exception is the "Infrastructure" alpha. It belongs to the "Resources" area of concern, so there is no similar-role alpha in the SEMAT model. The only exception is the "Infrastructure" set scorecards which is informed by the "Software System" SEMAT alpha.

Most times, the contents of the scorecards simply paraphrases those found in the SEMAT model. It is worth to note, however, that the original SEMAT model does not introduce a concept of prerequisite states – i.e. representation of required progress in one alpha to facilitate further progress in another alpha. In SEMAT, the states of different alphas reference each other only through conditions. In the Scorecard Deck, such cross-refence conditions are translated into prerequisites.

In cases when the contents of SEMAT and the Scorecard Deck match each other well, the fieldwork evidence – or other forms of additional evidence – is usually omitted. The evidence is however provided in cases of large deviations from the SEMAT model or when the fieldwork contains particularly illustrative examples of the importance of a particular condition or state.

A.1 Research Goals

This alpha is informed by the "Opportunity" alpha of the SEMAT model, which is defined as "the set of circumstances that makes it appropriate to develop or change a software system" (OMG, 2015, p. 27).

A.1.1.1 State 1, 2 and 3. Problem Identified, Research Goals Formulated, Research Goals Demanded

States 1, 2 and 3 (see Tables A.1, A.2 and A.3) are based on two first states of the respective alpha in the SEMAT model – "Identified" and "Solution Needed". While initially the states of two corresponding alphas had one-to-one mapping, "Problem Identified" has been since considered a separate state as a result of internal evaluation of the Scorecard Deck in the Dark Net study. That project brought a realisation that one particular problem (low effectiveness of law enforcement interventions into cybercriminal marketplaces) can lead to multiple research goals (understanding the factors of resilience of those marketplaces and assessment of changes in trade dynamics after disruptive events). As "problem" is defined in looser terms than "opportunity" in the SEMAT model, a separate condition that warrants *relevance of social data science* to the solution of the problem is added. The other conditions from the SEMAT model are slightly redistributed across the states in the Scorecards to better match the 3-state representation.

RESEARCH GOALS: PROBLEM IDENTIFIED	*A problem that can be addressed with the help of social data science is identified.*
Problem established	An unsolved problem is identified.
Social data science being relevant	There is a tentative idea of how a social data science project can contribute to solving this problem.

Table A.1: Research Goals: Problem Identified. Condition. No prerequisites specified.

A.1.1.2 State 4. Evaluated

This state (see Table A.4) is closely based off the "Opportunity: Value Established" state of the SEMAT model. It relaxes some of the strict conditions that SEMAT poses on the notion of value. In SEMAT, the value has to be "quantified either in absolute terms or in returns or savings per time period" (p. 30). This is not always applicable to social data science. As the fieldwork suggests, even the evaluation projects that the Open University hosted and that were to benefit a specific funding stakeholder did not aim to bring direct financial gain – rather, they could help a funder to make better investments in the future. Therefore, a more relaxed notion

RESEARCH GOALS: FORMULATED	*Potential goals of the social data science project are formulated.*
Stakeholders' needs established	The specific needs that motivate prioritising the identified research problem are identified.
Problem broken down	Where possible, the problem is broken down to its components and their potential roots.
Goals formulated	A list of goals that can be persued by a social data science project is compiled.
Stakeholders: Recognised	*Stakeholders and their potential modes of engagement are identified.*

Table A.2: Research Goals: Formulated. Conditions and prerequisites.

RESEARCH GOALS: DEMANDED	*The demand for social data science artefacts that addresses the goals is established.*
Key stakeholders interested	There is at least one stakeholder motivated to invest into further examination of the problem.
Artefacts needed	A need in social data science artefacts is confirmed.

Table A.3: Research Goals: Demanded. Condition. No prerequisites specified.

of a *benefit for key stakeholders* is introduced instead. Furthermore, the Scorecard Deck also introduces the concept on *wider impact*, i.e. social effects beyond the key/direct stakeholders. The Dark Net study informed this condition, as that study could have wider societal impacts that go beyond affecting the operations of the key stakeholders (law enforcement organisations). Finally, a condition on the SEMAT model that tied opportunity and requirements is translated into a prerequisite for the *research questions*, as the "Research Questions" alpha in the Scorecard Deck functionally corresponds with the "Requirements" alpha in the SEMAT model (see Section 4.1).

RESEARCH GOALS: EVALUATED	*The potential value of fulfilling the research goals is established.*
Value estimated	It is understood how the key stakeholders can benefit from fulfilment of the research.
Impact established	Wider societal impact of achieving the research goals is understood.
Success criteria established	The degree to which the research goals are achieved is measurable.
Research Questions: Outlined	*The research questions are formulated as a broad framework that sets the direction for the project.*

Table A.4: Research Goals: Evaluated. Conditions and prerequisites.

A.1.1.3 State 5. Viable

This state closely follows the equivalent state of the SEMAT model paraphrasing its conditions and translating some of them into prerequisites (see Table A.5). Prerequisites for achievement of early state for *"Team"*, *"Infrastucture"*, *"Data sources"* and *"Compliance"* alphas are added since having at least a rough idea about the first three is crucial for costing a social data science project and severe compliance issues may damage a project's viability.

RESEARCH GOALS: VIABLE	*The research goals can be achieved with the available resources.*
Constraints are acceptable	The resource- and other constraints put on the social data science endeavour allow for achieving the research goals.
Risks are manageable	The risks associated with progressing the endeavour are assessed and manageable to the degree necessary for the reseach endeavour.
Costs are allowable	The prospective costs of carrying out the endeavour are allowable.
Principle agreement achieved	All the team members and the stakeholder agree on the motivation behind the endeavour.
Viability established	Carrying out the research endeavour is clearly viable.
Artefacts: Envisioned	*There is a clear, valid and agreed vision of the final artefacts.*
Stakeholders: In Agreement	*The stakeholder representatives are in agreement.*
Team: Seeded	*There are initial (seed) team members and the mechanisms of expanding the team.*
Infrastructure: Architecture Selected	*Architecture has been selected. It addresses the key technical risks and any applicable organisational constraints.*
Data Sources: Identified	*The possible relevant data sources are identified and a preliminary list of appropriate data sources is compiled.*
Compliance: Considered	*It is identified which ethical and legal concerns the project raises.*

Table A.5: Research Goals: Viable. Conditions and prerequisites.

A.1.1.4 State 6. Addressed

This state closely follows the equivalent state of the SEMAT model (see Table A.6). The prerequisites added on top simply reflect the states of other alphas that are associated with addressing research goals.

RESEARCH GOALS: ADDRESSED	*The research undertaken within the endeavour addresses the goals.*
Contributions towards the goals secured	The research has brought some outcomes that evidently contribute towards the goals.
Value seen	The intermediate research outcomes are valuable, as agreed by the team and the stakeholders.
Initial use facilitated	Stakeholders can at least have a start at using the reseach outcomes for their benefit.
Research Questions: Answered	*Through applying the analytical methods to the acquired- and quality assured data, actionable answers to the research questions are derived.*
Artefacts: Iterated	*The research methods are executed and inform an iteration of artefacts with the research findings.*
Team: Performing	*The team is efficient and effective at progressing its work.*

Table A.6: Research Goals: Addressed. Conditions and prerequisites.

A.1.1.5 State 7. Fulfilled

This state closely follows the "Opportunity: Benefit Accrued" state of the SEMAT model, although (as is the case with the *State 4. Evaluated*), the SEMAT's language of financial returns and gains is relaxed (see Table A.7). The prerequisites play the same role as with the previous state.

RESEARCH GOALS: FULFILLED	*The produced artefacts fulfil the research goals.*
Artefacts in working order	The released research artefacts are of desired quality.
Fulfilment criteria met	The fulfilment criteria are met for each of the research goals.
Benefit emerging	There is evidence of the a sustained benefit for the endeavour's stakeholders.
Research Questions: Utilised	*Answers to the research questions are in use.*
Artefacts: Released	*Research artefacts are produced and presented to the stakeholders.*
Stakeholders: Satisfied with Artefacts	*The expectations of the stakeholder representatives have been achieved.*

Table A.7: Research Goals: Fulfilled. Conditions and prerequisites.

A.2 Stakeholders

This alpha is informed by the equivalent alpha of the SEMAT model, which is defined as "the people, groups, or organizations who affect or are affected by a software system" (OMG, 2015, p. 23).

A.2.1.1 State 1. Recognised

This state closely follows the equivalent state of the SEMAT model (see Table A.8). Additionally, recognition of *tacit stakeholders* is acknowledged. Tacit stakeholders are individuals and groups who do not directly interact with the project, its findings and artefacts, but are affected by extension. In the Dark Net study, some of those tacit stakeholders are drug users and their families. To an extent, the distributors of illicit goods were also tacit stakeholders – this illustrates the point that an effect that a social data science project can have on its stakeholders does not necessarily need to be positive.

Human participants are also recognised as a specific type of stakeholder group. Indeed, if a study involves participants, their interests and concerns have to factored in at the very least for the ethical reasons. The infamous study of emotional contagion on Facebook (Kramer et al., 2014) is an example of *no* consideration for participants is stakeholders – and the backlash it caused was massive.

It is worth noting that, since social data science projects often have wide impact, involvement (and even identification) of all stakeholder groups is unfortunately sometimes impractical or impossible. For example, in the case of the Hit List case study, our team had limited access to audience research data and no resources to conduct audience surveying ourselves. For this reason, the quality of work was mainly assessed through the professional judgement of the producers involved and through examining occasional feedback left by social media users. While the show's team did put much effort and reflection into how the show's production (and the social data science behind it) were approached, it is impossible to eliminate the probability that systematic audience research could have shaped the process.

A.2.1.2 State 2, 3 and 4. Represented, Involved and In Agreement

These states closely follow the equivalent states of the SEMAT model while adding conditions in relation to the aforementioned *tacit stakeholders* (see Tables A.9, A.10 and A.11). The condition of *stakeholders supporting ways of working* is shifted from State 2 to State 4. The fieldwork shows that the ways of working develop throughout the course of the project. For example, Section 3.2.7.2 shows how the team collaboration modes naturally evolved with the

STAKEHOLDERS: RECOGNISED	*Stakeholders and their potential modes of engagement are identified.*
Groups identified	All the different groups of stakeholders that are, or will be, affected by the research project are identified, including tacit stakeholders.
Involvement system established	It is known which stakeholder groups will actually have representatives specifically appointed to interact with the project and its team. At least the funders and the beneficiaries are considered.
Human participants considered	If the research involves human participants, those are considered stakeholders represented via participation in the research activities.
Roles defined	The roles that representatives of each stakeholder group should play in their interactions with the project team (e.g. superintendence, consultancy, assessment) are defined.
Research Goals: Problem Identified	*A problem that can be addressed with the help of social data science is identified.*

Table A.8: Stakeholders: Recognised. Conditions and prerequisites.

team's understanding of how complex it was to ensure inter-coder reliability for some of the data annotation variables. Thus, it may be impossible for the stakeholder representative to make an informed judgement about the team's ways of working early on in the project.

In regard to State 3, the fieldwork illustrates the difference between *stakeholder feedback* and *stakeholder engagement*. The British Council provided sufficient feedback on interim results of the Shakespeare Lives evaluation project. This allowed to progress such areas of the project as social media analysis, since the project team could plan the early stages of work basing on their research experience, so that the British Council's feedback could be used to steer the direction of research rather than to provide fully detailed requirements. Yet, lack of the British Council's engagement beyond giving feedback led to a skewed perception of their priorities over some illustrative data visualisations, most notable the interactive calendar of the Shakespeare Lives programme (see Section 3.2.3.3).

A.2.1.3 State 5. Satisfied with Progress

The state corresponds to the "Stakeholders: Satisfied for Deployment" state of the SEMAT model (see Table A.12). The second condition in this state is more relaxed than the original SEMAT condition "The stakeholder representatives confirm that they agree that the system is ready for deployment". This is done to represent the iterative nature of work on artefacts and stakeholder feedback on them. The fieldwork experience of the Open University based evaluation projects

STAKEHOLDERS: REPRESENTED	*The mechanisms of stakeholder involvement and the stakeholder representatives are known.*
Responsibilities agreed	The stakeholder representatives are happy with the assigned responsibilities.
Power warranted	If required, the stakeholder representatives have been warranted the power to act in accordance with their responsibilities.
Collaboration approach agreed	The stakeholder representatives have agreed on the ways of collaborating.
Tacit stakeholders accounted	If there are tacit stakeholder groups who cannot be represented, it is decided how their needs and aspirations may be accounted for.

Table A.9: Stakeholders: Represented. Condition. No prerequisites specified.

STAKEHOLDERS: INVOLVED	*The stakeholder representatives are contributing towards their responsibilities through engagement in the project.*
Stakeholder assistance in place	The stakeholder representatives provide the agreed support to the project team.
Stakeholder feedback involved	The team gets feedback and input to key decisions from the stakeholder representatives as and when required.
Changes communicated	Should the circumstances of a stakeholder group change, those changes are swiftly communicated by the stakeholder representatives.

Table A.10: Stakeholders: Involved. Condition. No prerequisites specified.

strongly suggests that, as the project reports went through many drafts before being accepted.

A.2.1.4 State 6. Satisfied with Artefacts

The state corresponds to the "Stakeholders: Satisfied in Use" state of the SEMAT model (see Table A.13). A condition of *wider groups involvement* is added. In social data science, depending on the type of an artefact, it may be hard to judge how widely the artefact is used among the stakeholders – for example, it is not necessarily clear what it means to use a report or a publication. Therefore, making an extra check for this condition may be worthwhile. In addition, the last condition is relaxed from explicit positive feedback from stakeholders to some sort of *evidence of stakeholder satisfaction*. This is again mostly caused by tacit stakeholders and other stakeholder groups who are not necessarily directly approachable, but whose impression of the project artefacts is of importance.

STAKEHOLDERS: IN AGREEMENT	*The stakeholder representatives are happy with the project setup and its ongoing state.*
Minimal expectations agreed	The stakeholder representatives are happy to agree upon the minimal desired research outcomes.
Satisfaction with involvement achieved	The stakeholder representatives feel good about their degree of involvement.
Agreement with the ways of working achieved	The stakeholder representatives adapt their involvement to the ways of working employed be the team.
Mutual respect achieved	The stakeholder representatives and the team treat each others' involvement and input with respect.
Balance agreed	The contradictory interests and perspective of different stakeholder groups are balanced in a manner that the stakeholder representatives are happy to agree on.
Risk of tacit stakeholder discrimination managed	It is made sure that the risk of discriminating against the tacit stakeholders is accounted for and kept within allowable margins.
Artefacts: Outlined	*There is a tentative understanding of the possible forms of artefacts and the motivations behind them.*
Ways of Working: Informed	*The principles, policies, tools and practices that inform, facilitate and shape the way of working are selected.*
Team: Formed	*There is a sufficient number of engaged team members to progress towards the team's goals.*
Work: Piloted	*The very first steps of work undertaken to better understand the scope of further work and the involved complexities.*

Table A.11: Stakeholders: In Agreement. Conditions and prerequisites.

A.3 Infrastructure

The scorecards for this alpha are informed by the scorecards for the "Software System" alpha of the SEMAT model, which is defined as a "system made up of software, hardware, and data that provides its primary value by the execution of the software" (OMG, 2015, p. 38). As seen, this definition is inclusive of all infrastructure components – not only software per se – and thus is directly applicable to the "Infrasturcture" alpha. This is not to contradict Section 4.1.2, which claims that the SEMAT "Software System" alpha corresponds to "Artefacts" in the Essence of Social Data Science. The correspondence stated in Section 4.1.2 is in terms of the role that these alphas play in their respective types of projects. Indeed, a software system aims to satisfy the project requirements and serve the stakeholder needs in a software engineering project; similarly, research artefacts aim to contain answers to the research questions and be of value for the stakeholders in a social data science project. By contrast, the match between the Scorecard Deck's "Infrastructure" scorecards and the SEMAT's "Software System" scorecards

STAKEHOLDERS: SATISFIED WITH PROGRESS	*The stakeholder representatives see the progress towards their needs.*
Feedback on artefact iterations provided	The stakeholder representatives assess the artefact iterations and provide their feedback in accordance with the interest of their stakeholder group.
Satisfaction with progress confirmed	The stakeholder representatives agree that the artefact iterations are progressing in matching the stakeholders' needs.
Artefacts: Iterated	*The research methods are executed and inform an iteration of artefacts with the research findings.*

Table A.12: Stakeholders: Satisfied with Progress. Conditions and prerequisites.

STAKEHOLDERS: SATISFIED WITH ARTEFACTS	*The expectations of the stakeholder representatives have been achieved.*
Artefacts accepted	The stakeholder representatives accept the artefacts.
Wider groups involved	The artefacts have been tried by wider stakeholder groups rather than only be the representatives.
Wider satisfaction confirmed	There is evidence that the stakeholder groups as whole are satisfied with the artefacts.
Artefacts: Released	*Research artefacts are produced and presented to the stakeholders.*

Table A.13: Stakeholders: Satisfied with Artefacts. Conditions and prerequisites.

is in terms of the *specific real-world objects* that these scorecards are concerned with – and, crucially, of the states that these objects go through within the course of a project.

The SEMAT model provides some of its most rigorous and detailed advice to management of specifically the "Software System" alpha. This is not surprising, since the model is developed by software engineers and for software engineering teams. The Essence of Social Data Science model retains this approach. While a social data science project can occasionally do with little software engineering in its 'traditional' form of designing and implementing own code, they extensively rely on what Sommerville calls "construction by configuration" (Sommerville, 2008) – building a system that satisfies the needs of the project by reuse of software components (external and previously-built internal). The emphasis on reuse in construction by configuration suggests that employing proper software engineering practices is no less important than in traditional software engineering. In social data science, the benefits of reuse are quite self-evident when the reused software components facilitate some form of advanced data processing functionality. However, the experience of the studied project suggests that even in case of more trivial data

operations, *reuse can bring research advantages that go beyond saving human effort* for more creative tasks. For example, in the InfoMigrants projects the data visualisation library that was originally built during the Shakespeare Lives evaluation (see Section 3.2.6.4) was reused. The reuse of the library led to (a) better stakeholder communication due to more effective final editing of reports in the context of approaching deadlines, and (b) better *research flow* when data visualisations were used for exploratory analysis during the rare (and thus especially valuable) brainstorming team sessions: high quality charts produced "on the spot" proved to be a powerful tool for generating and testing idea.

For the reasons outlined above, the "Infrastructure" alpha in the Scorecard Deck close follows its SEMAT prototype and all its the states correspond to their homonyms in SEMAT's "Software System" alpha. Yet, there are some deviations caused by a supportive role infrastructure plays in the social data science project (instead of being the direct artefact of such project) and the project scope of the alpha. Below these deviations are discussed in detail.

A.3.1.1 State 1. Architecture Selected

Compared to the SEMAT prototype, a condition for defining boundaries of the software system is removed (see Table A.14). This condition is only relevant when a software system by itself is the promised outcome of a project, as in such cases the level of the commitment has to be scoped. This is indeed the case in software engineering projects; however, in social data science infrastructure plays a supporting role and addresses the needs of a team – not of project stakeholders. Furthermore, *hardware platforms* are split into readily available on the on hand and relevant and potentially available on the other. The fieldwork suggests that the hardware component of infrastructure in a social data science project is often assembled ad-hoc with a priority given to existing pieces of hardware even if they are not best-suited for the work. For example, the computations for the Dark Net study were made on a server that just happened to be in possession on one of the project investigators and otherwise stood idle.

A.3.1.2 State 2. Demonstrable

The scorecard closely follows its SEMAT prototype, although the conditions again stress that a project infrastructure should satisfy the needs of its teem – not of its stakeholders (see Table A.15).

A.3.1.3 States 3 and 4. Usable and Ready

Since infrastructure is not by itself an artefact of a social data science project, the need to understand its *value* (as suggested as a condition for the "Software System: Usable" state of the

INFRASTRUCTURE: ARCHITECTURE SELECTED	*Architecture that suits the needs of a social data science projects has been selected.*
Circumstances considered	Technological needs determined by the suggested data sources, analytical methods and team / work characteristics are considered.
Criteria agreed	The criteria that can be used as a basis for determining which architecture to select are agreed on.
Hardware platforms identified	The readily available and relevant potentially available hardware platforms are identified.
Technologies selected	Technologies to be employed are selected. This includes programming languages if programming component is involved.
Infrastructure organisation outlined	Key decisions on how the selected technologies will be integrated are made.
Buy, build and reuse decided	The decisions on which components to purchase, which to reuse from previous work and which to build are made.
Risks agreed	There is understanding of key risks associated with the infrastructure. The risks are acceptable.
Data Sources: Identified	*The possible relevant data sources are identified and a preliminary list of appropriate data sources is compiled.*
Analysis Methods: Selected	*Candidate analysis methods are identified and selected.*
Team: Seeded	*There are initial (seed) team members and the mechanisms of expanding the team.*
Work: Initiated	*There is a serious initiative to start the work.*

Table A.14: Infrastructure: Architecture Selected. Conditions and prerequisites.

SEMAT model) can be relaxed – instead, understanding the *appropriateness* of the infrastructure is suggested. For similar reasons, the SEMAT's condition of desiring to use a software system is not reflected in the scorecard of the "Infrastructure: Ready" state at all, as using infrastructure is not voluntary in a social data science project (see Table A.16).

A.3.1.4 State 5. Operational

This state closely follows its SEMAT prototype, however it puts a strict condition of *fully supporting a social data science endeavour* (see Table A.18). Indeed, here the infrastructure of a whole project is implied, while a software system can be scoped arbitrarily.

INFRASTRUCTURE: DEMONSTRABLE	*A prototype executable version of the infrastructure is available. It demonstrates the fit of the architecture for the project purposes and for required testing.*
Architecture demonstrated	The prototype architecture demonstrates its key characteristics.
Performance being measurable	Infrastructure can be test-run and its performance is measurable.
Hardware shown	Core hardware configurations can be observed.
Interfaces shown	Core interfaces can be observed.
Integration demonstrated	If relevant for the needs of the research project, it is known how the infrastructure can be integrated with those of other projects and of the wider organisation.
Initial agreement achieved	The project team considers the infrastructure to be appropriate for their needs.

Table A.15: Infrastructure: Demonstrable. Condition. No prerequisites specified.

INFRASTRUCTURE: USABLE	*The core components of the infrastructure are in place and can be used.*
Operability achieved	The team members who use the infrastructure can operate it.
Functionality tested	The infrastructure's functionality is successfully tested.
Performance accepted	The infrastructure's performnace satisfies the research needs.
Defect levels accepted	The frequency and severity of defects are acceptable for the research needs.
Documentation completed	The documentation of the infrastructure is complete and comprehensive.
Appropriateness being clear	The way the infrastructure can successfully support execution of the research is clear.

Table A.16: Infrastructure: Usable. Condition. No prerequisites specified.

A.3.1.5 State 6. Retired

Since the "Infrastructure" alpha – as all alphas in the Social Data Science Scorecard Deck – is scoped to the level of an individual project, additional conditions on *satisfaction of project needs* and on *possibility of reuse* are put onto its "Retired" state compared to its prototype in the SEMAT model (see Table A.18). Ensuring *Recomputability* is listed as an additional condition as well. It is inspired by the "Recomputation manifesto" (Gent, 2013) that states that sometimes preserving research workflows is not enough to rerun the data processing jobs. This is confirmed by Glatard et al. (2015) who show that even minor differences in system configuration may cause reruns to fail. Gent (2013) therefore suggests to strive to "recomputability" instead – i.e. for capturing the whole infrastructure in a form of a virtual machine thus creating a virtual copy of

INFRASTRUCTURE: READY	*The infrastructure is acceptable to be continuously deployed in the project.*
Documentation available	Documentation for installation, use and maintenance of the infrastructure are available.
Infrastructure accepted	There are no immediate concerns about the infrastructure that would prevent its deployment.
Support practices in place	It is clear who and how will support the use of infrastructure on the operational level.

Table A.17: Infrastructure: Ready. Condition. No prerequisites specified.

INFRASTRUCTURE: OPERATIONAL	*The infrastructure is in continuous operational use.*
Team satisfied	The team is able to use the infrastructure for the project needs.
Endeavour resourced	The infrastructure provides all the relevant support for progressing the endeavour.
Technical support in place	The agreed operational support for the infrastructure is fully in place.

Table A.18: Infrastructure: Operational. Condition. No prerequisites specified.

both hardware and software systems.

INFRASTRUCTURE: RETIRED	*The support for the infrastructure is no longer provided.*
Project needs satisfied	All the research- and other activities of the project that require technical infrastructure are completed.
Infrastructure discontinued	The use of the infrastructure has been discontinued. If relevant, the infrastructure is dissolved.
Support stopped	Support of the infrastructure is no longer maintained.
Possibility of re-use assured	If possible, there is an established way to reinvoke the infrastructure in future endeavours.
Recomputability assured	If possible and relevant, the full snapshot of the infrastructure is stored as a virtual machine that can be used to rerun the project computations.
Work: Concluded	*The team has finished its work on the social data science endeavour.*

Table A.19: Infrastructure: Retired. Conditions and prerequisites.

A.4 Team

This alpha is informed by the equivalent alpha of the SEMAT model, which is defined as "a group of people actively engaged in the development, maintenance, delivery, or support of a specific software system." (OMG, 2015, p. 48). The alpha follows its prototype almost verbatim and the progression of states is exactly the same in both models. Such close match is possible since issues of team management transcend those of specifically software engineering or social data science and are equally applicable to any kind of project. This manifests, for example, in existence of widely recognised field-agnostic team leadership guidelines, best practices and theories (e.g. Kozlowski et al., 1996; Zaccaro et al., 2001). The minor discrepancies that still exist between the Scorecard Deck's and the SEMAT's representation of the "Team" alpha are discussed below.

A.4.1.1 State 1. Seeded

The scorecard closely follows its SEMAT prototype, but further specifies a couple of conditions. (see Table A.20). The condition of understanding the constraints on the team work is reformulated into understanding *the principal mode of team collaboration*, i.e. whether the team will be co-located or work remotely. The experience of all studied projects suggest that the issue of co-location is of crucial importance and transcends the issues of ways of working. In the Shakespeare Lives project, the choice of Sina Weibo as the primary source of Chinese data was partially caused by the difficulties of collaboration on Twitter data acquisition criteria between a project management located in England and a researcher located in China (see Section 3.2.7.1). In the Hit List case study, my role of data consultant for the production team was motivated by the fact that the analyst worked remotely and could not be that easily questioned above the particularities of weekly data in real-time (see Section 3.3.2).

Another significant change is that the condition on understanding *team commitment* is clarified in terms of time and effort each team member commits to put. In all the studied project, the team members had other professional or study responsibilities. For example, most analysts in the Open University's evaluation projects were either PhD students or early career researchers doing consultancy work, the data science team (including the lead analyst) on the Hit List projects were all academics in British Universities and the most producers were freelancers.

A.4.1.2 State 2. Formed

The state follows its SEMAT prototype with little to no deviations (see Table A.21).

TEAM: SEEDED	*There are initial (seed) team members and the mechanisms of expanding the team.*
Seed members ready	There are initial team members who serve as a backbone for the team and who are committed to the social data science project.
Principle mode of team work considered	It is known whether the team will be co-located or collaborate remotely, and which degree of collaboration is expected. The specific ways in which this collaboration will be implemented (i.e. ways of working) do not need to be known.
Growth mechanisms set	Procedures and practices required to recruit team members are defined.
Desired backgrounds identified	It is known what broad competencies are required within the team and how they map to professional backgrounds of potential team members.
Team's responsibilities outlined	The responsibilities of the team as a whole are broadly outlined.
Commitment requirements clarified	It is clear how much time and effort the prospective team members are expected to put into work throughout the course of the project.
Preferred size set	The preferred team size is defined within a reasonable margin of error.
Governance practices established	The relationships between the key governing stakeholders and the project team are defined.
Team management model chosen	The organisational structure and the practices of team leadership are chosen.
Research Goals: Formulated	*Potential goals of the social data science project are formulated.*
Work: Initiated	*There is a serious initiative to start the work.*

Table A.20: Team: Seeded. Conditions and prerequisites.

A.4.1.3 State 3. Collaborating

This states adds an additional condition of *awareness of wider team progress* compared its the SEMAT prototype (see Table A.22). In bigger social data science projects, it is unfortunately not uncommon for sub-teams that work on different strands to effectively lose touch with each other. This effectively happened at one moment of time in the Shakespeare Lives project, as the work on the Cultural Value strand was isolated from the work on the strand studying digital engagement (which I participated in) until later on in the project. A more collaborative approach could have potenrially made interpreting the findings of the digital engagement strand in the terms of the Cultural Value Model a bit easier (see Section 3.2.5.3).

TEAM: FORMED	*There is a sufficient number of engaged team members to progress towards the team's goals.*
Team size being sufficient	The number of team members allows for progressing the endeavour.
Responsibilities set	Each team member knows and understands responsibilities that are assigned to them and to the other team members.
Work accepted	Each team member knows and accepts the work they need to do to fulfil their responsibilities.
Competencies matched	Each team member can efficiently use their competences to complete their share of work.
Collaboration seeded	At least one team meeting have been held and introduction of team members have happened.
External parties identified	All external parties who contribute to the work are identified. The external parties may or may not be part of stakeholder groups.
Communication established	The mechanisms for team members' communication are established. Those include both modes of communication and, if appropriate, schedule / frequency of communication.
Commitment achieved	An appropriate form of commitment to the project (e.g. a work contract) is secured for each team member.
Work: Piloted	*The very first steps of work undertaken to better understand the scope of further work and the involved complexities.*
Ways of Working: Informed	*The principles, policies, tools and practices that inform, facilitate and shape the way of working are selected.*

Table A.21: Team: Formed. Conditions and prerequisites.

A.4.1.4 State 4. Performing

The only significant deviation from the SEMAT prototype that this state's scorecard bears is the absence of a condition on eliminating wasted work (see Table A.23). In social data science, as in research in general, the work that is "wasted" in the sense that it does not contribute to the final artefacts or do not add much value to them may still inform better, refined questions or methods. In this sense, it is actually useful. The aforementioned example of switching to more qualitative analysis in the InfoMigrants project in its second year after seeing low added value of some previously taken strands of research is a great example of that (see Section 4.3.4.4).

A.4.1.5 State 5. Adjourned

The state follows its SEMAT prototype (see Table A.24). It omits the condition that team members should be free to move on work on other projects, since, as mentioned above (see Section A.4.1.1), in social data science team members are often free to work on several projects

TEAM: COLLABORATING	*The members of the team effectively collaborate with each other.*
Mutual knowledge achieved	The team members have acquired understanding of each others' personal and professional qualities sufficiently for effective collaboration.
Team united	The work is performed in cohesive cooperation. There are little to no frictions between the team members.
Communication established	The team members communicate with their colleagues honestly, openly and with ease.
Internal awareness established	Each team member has at least an idea about the progress of all other team members, including those with whom they do not need to directly collaborate.
Team being focused	The team works towards achieving the research goals and producing the best possible artefacts
Ways of Working: Employed	*The conditions required for the work to start are met.*

Table A.22: Team: Collaborating. Conditions and prerequisites.

TEAM: PERFORMING	*The team is efficient and effective at progressing its work.*
Work routinely progressed	The team efficiently progresses the required work and achieves the planned milestones.
Adaptivity achieved	The team effectively manages to adapt to changes in external circumstances.
Self-sufficiency achieved	The team does not require external interventions to address any rising issues.
Performance being efficient	The team's efforts are helpful for achieving the project's goals. Any effort wastes are due to unavoidable uncertainties of social data science, not due to inefficiencies.
Ways of Working: Optimised	*The ways of working is executed fluently and efficiently with minimal costs of time and effort spent on non-productive activities.*

Table A.23: Team: Performing. Conditions and prerequisites.

simultaneously anyway.

A.5 Work

This alpha is informed by the equivalent alpha of the SEMAT model, which is defined as "activity involving mental or physical effort done in order to achieve a result" (OMG, 2015, p. 52). The scorecards for this alpha deviate from their prototype more than for the above discussed "Team"

TEAM: ADJOURNED	*The team has finished its work on the social data science project.*
Commitments fulfilled	The team has finished implementing its commitments.
Members dismissed	The team members are free from commitments on the social data science project.
Work: Concluded	*The team has finished its work on the social data science endeavour.*
Ways of Working: Preserved	*The new and updated practices, policies and managerial tools are preserved for further use in future social data science projects.*

Table A.24: Team: Adjourned. Conditions and prerequisites.

alpha, as they aim to reflect the higher level of uncertainty involved in doing research.

A.5.1.1 State 1. Initiated

The state is informed by its homonym in the SEMAT model, although it removes the condition of understanding the priorities in work (see Table A.25). In the studied projects the priorities shifted quite significantly and evolved naturally. For example, in the Dark Net study, it seemed initially that the focus would be on developing data acquisition code for as many marketplaces as possible and on creating visual representations of the data. However, the work was soon re-prioritised to building tools that would support analysis of the acquired data, as such analysis was not trivial while even evidence from one platform was already very rich.

WORK: INITIATED	*There is a serious initiative to start the work.*
Body of work defined	Outline of the project goals and artefacts informs the principal body of work required.
Scope outlined	There is a preliminary understanding of the scope of the work in terms of required time, human effort and resources.
Stakeholder involvment in work identified	It is clear which stakeholder group initiates the work on the social data science endeavour, which funds it and which assesses the resulting artefacts.
Artefacts: Outlined	*There is a tentative understanding of the possible forms of artefacts and the motivations behind them.*
Research Goals: Demanded	*The demand for social data science artefacts that addresses the goals is established.*
Stakeholders: Represented	*The mechanisms of stakeholder involvement and the stakeholder representatives are known.*

Table A.25: Work: Initiated. Conditions and prerequisites.

A.5.1.2 State 2. Piloted

This state is absent in the SEMAT model and stems from the fieldwork observation (see Table A.26). All studied project have included a pilot stage in some shape of form. In the earliest weeks of the Hit List show only Twitter data were acquired and analysed, as this was the platform with which the data science team had had past experience. Restricting the work to Twitter analysis allowed to get a taste for the *complexity* of issues associated with the work – both methodological such as separating news from noise and topics from each other (see Section 3.3.4) and organisation such as communicating remotely with the production team during their meetings – and yet to do that in a "lite mode", i.e. without needing to solve the hard task of aggregating evidence from multiple platforms.

In the Dark Net study, the pilot work involved developing an early version of scraping code for a particular marketplace for a limited-scope test scrape. This allowed to identify and scope the *risks* associated with data scraping – interfering captchas, low uptime of markets, slowness of Tor network and others. By the time the core work on the project started, the marketplace for which the pilot code had been developed was shut down by law enforcement and thus the code could not be reused. However, the lessons from the pilot work was still valuable not only because the identified risks did translate to other marketplaces, but also because it motivated us to *consider the compliance issues* early on in the project and thus have more time to formulate a coherent compliance strategy for the main body of work.

As seen in both examples above, the pilot work was not trivial. It involved dealing with a subset of the same tasks that the main body of work would subsequently include. Thus, it could be considered *taking first steps* and it did require *enabling* – at least in the way of putting together *preliminary infrastructure* and having *some team members available*.

A.5.1.3 State 3. Prepared

The state is informed by its homonym in the SEMAT model, but with many adjustments. Some of the conditions (such as understanding risk exposure) are removed because they are already covered by the introduced "Work: Piloted" state (see above). The previously omitted condition of understanding work priorities (see Section A.5.1.1) is returned here since the pilot work may allow to set those (see Table A.27). Finally, some are reformulated in terms of prerequisites – for example, "Acceptance criteria are defined and agreed with client." (p. 55) is changed to "Artefacts: Envisioned", since a valid vision for artefacts already includes the requirements for those (see Section 4.3.5.2).

WORK: PILOTED	*The very first steps of work undertaken to better understand the scope of further work and the involved complexities.*
First steps enabled	There is a good understanding of the initial chunks of work that the seed team members could and should perform.
First steps taken	The seed team members have made an attempt at the initial chunks of work.
Complexity understood	The initial steps of work further inform the complexity of the work ahead.
Risks known	There is an understanding of the risks that may hinder the work. As far as possible, the magnitude of adverse consequences and the chance of happening are estimated for these risks.
Effort re-scoped	The scope of the effort required for the work is better understood and put into correspondence with what is practically available.
Infrastructure: Demonstrable	*A prototype executable version of the infrastructure is available. It demonstrates the fit of the architecture for the project purposes and for required testing.*
Team: Seeded	*There are initial (seed) team members and the mechanisms of expanding the team.*
Compliance: Considered	*It is identified which ethical and legal concerns the project raises.*

Table A.26: Work: Piloted. Conditions and prerequisites.

A.5.1.4 State 4. Started

This state follows its SEMAT prototype and adds an extra condition of *documenting the work* (see Table A.28). The importance of documenting the work has been shown above several time – be that documentation for the compliance procedures (see Section 4.3.6.3) or for data provenance (see Section 4.3.1.3). Also, the requirement of merely monitoring the work replaced with a stronger one on *iterating* on the work's intermediate outcomes. The importance of critical reflection and iteration on work outcomes can be seen in the context of Hit List work, as over time the analyst and the production team were getting progressively better in knowing the idiosyncrasies of the collected data and could handle them more efficiently (see Sections 3.3.3.3 and 3.3.5.1).

A.5.1.5 State 5. Under Control

The SEMAT prototype is closely followed by this state (see Table A.29). Some conditions to the quality of performance are omitted as they are effectively captured by the "Team: Perfroming" prerequisite. Also, while the SEMAT model suggests having "measure" (p. 55) for progress,

WORK: PREPARED	*The conditions required for the work to start are met.*
Funding ready	The funds are ready to be disposed on the work.
Priorities set	If the project involves multiple questions to answer and / or artefacts to produce, those are prioritised.
Overall plan established	There is a feasible plan for the work to be undertaken, which is still sufficiently flexible to account for the uncertainties associated with research work.
Infrastructure: Ready	*The infrastructure is acceptable to be continuously deployed in the project.*
Team: Formed	*There is a sufficient number of engaged team members to progress towards the team's goals.*
Artefacts: Envisioned	*There is a clear, valid and agreed vision of the final artefacts.*
Ways of Working: Informed	*The principles, policies, tools and practices that inform, facilitate and shape the way of working are selected.*
Compliance: Progressed	*The team has taken the steps necessary to secure the compliance resources and is ready to actively engage in the main bulk of research subject to secured ethical clearances.*

Table A.27: Work: Prepared. Conditions and prerequisites.

here the condition is relaxed to qualitative *understanding of progress* to reflect the open nature of research work.

A.5.1.6 State 6. Concluded

The state strictly follows its SEMAT prototype (see Table A.30).

A.6 Ways Of Working

This alpha is informed by the "Way-of-Working" alpha of the SEMAT model, which is defined as "the tailored set of practices and tools used by a team to guide and support their work."

A.6.1.1 State 1. Informed

The scorecard for this state is heavily informed by the "Way-of-Working: Foundation Established" state in the SEMAT model, but it suggests more possible sources to inform ways of working (see Table A.31). They are analogous for those that can be used to inform compliance practices (see Sections 4.3.6.1 and 4.3.6.2). One of the conditions implicitly suggests having a *repository for past practices*. This does not reflect any particular fieldwork finding, but rather is a suggestion that reflects one of the assumed uses of the Social Data Science Scorecard Deck, as different

WORK: STARTED	*The team has started to put its required efforts into achieving the project's goals.*
Research started	The work requires to answer the research questions is started.
Preliminary findings iterated	The initial findings start to emerge. Even if not being directly useful for the artefacts by themselves, they at least provide a direction for further work.
Work broken into chunks	Possibly with the help of interim findings, the work starts to get iteratively broken into manageable and actionable individual pieces.
Progress tracked	The progress of the work is routinely tracked and assessed.
Team members engaged	The members of the team take on the appearing tasks and progress them.
Documentation initiated	The team has started to document the research- and overall work process to a degree that at least allows traceability of the work's results and appropriate level of accountability towards stakeholders.
Compliance: Secured	*The compliance resources required for the project are secured to the degree that the team can fully engage in the main body of the research work.*

Table A.28: Work: Started. Conditions and prerequisites.

alphas and states in the deck can be used as a glossary to classify emerging practices, thus naturally giving a convenient way to store and navigate them.

A.6.1.2 State 2. Employed

The scorecard for this state is heavily informed by the "Way-of-Working: In Use" state in the SEMAT model, although it also suggests to *note discrepancies* between the ways of working and the involved work (see Table A.32). One example of such discrepancis could be found in the Shakespeare Lives project. In that project, a project management platform TeamWork was employed to strengthen team collaboration (see Section 3.2.7). The platform was immensely useful, however at the early stages of the project it did require the team members to adjust their practices. For example, it took the team members a bit of time to actually switch to posting on TeamWork instead of writing emails. What took even longer was developing habits on inform relevant people within the team about new posts. When a researcher posted something on TeamWork, they could choose whom to send an email notification to. At the beginning of using TeamWork there was a strong tendency to notify an excessive number of people, since including all was the default but over time we developed unwritten laws on whom from the project managers and investigators to notify about each type of posting.

WORK: UNDER CONTROL	*The work goes productively and in a risk-managed environment.*
Tasks handled	There is a routine progress over the core tasks of the project. The research questions are getting their answers.
Uncertainties managed	There is a working system of practices oriented at managing risks and uncertainties that, to an acceptable level of confidence, does not allow unforeseen circumstances to prevent further work.
Expectations managed	The requirements for the desired artefacts and to the research findings are revised to better match the practicalities of the involved work.
Time constraints are held	The work progresses with a tempo within the allowed time constraints.
Progress firmly understood	There is a robust system of progress tracking that allows the team to firmly understand the degree of work progress and to communicate it to the stakeholders.
Documentation managed	The team makes sure that the documentation is constantly updated and satisfies the requirements of accountability and provenance traceability.
Infrastructure: Operational	*The infrastructure is in continuous operational use.*
Team: Performing	*The team is efficient and effective at progressing its work.*
Compliance: Maintained	*The team works on the project while reactively and proactively maintaining compliance.*

Table A.29: Work: Under Control. Conditions and prerequisites.

A.6.1.3 State 3. Adapted

The scorecard follows the "Way-of-Working: In Place" SEMAT state and, in line with the adjustments to the previous state, adds conditions on *making corrections* to the employed ways of working to better fit the work and *seeing improvements* to the work (see Table A.33). In the aforementioned example of using TeamWork for communication and notifying relevant team members about new posts, over time we got progressively less communication were someone expressed being annoyed with too much notifications. Simultaneously, useful brainstorming discussion threads where whole sub-teams participated started to appear (such as the one on intercoder reliability, see Section 3.2.7.2).

A.6.1.4 State 4. Optimised

The state follows the "Way-of-Working: Working well" SEMAT state and adds a condition on *minimising overheads* caused by the ways of working (see Table A.34). An example of such

WORK: CONCLUDED	*The team has finished its work on the social data science endeavour.*
Tasks fulfilled	The tasks of the work have been fulfilled and their outcomes can be demonstrated.
Work accepted	The key stakeholders agree that the work may be considered finished and accept its outcomes.
Archiving completed	The results of the work, the documentation and any supporting evidence are properly achieved to the degree allowed by the compliance requirements.
Stakeholders: Satisfied with Artefacts	*The expectations of the stakeholder representatives have been achieved.*
Research Goals: Fulfilled	*The produced artefacts fulfil the research goals.*

Table A.30: Work: Concluded. Conditions and prerequisites.

minimisation was my inclusion onto the Hit List team as a data consultant (see Section 3.3.2), which allowed for more rapid investigations into the data if such an investigation was required by the production team and for more seamless dialogue about the data.

A.6.1.5 State 5. Preserved

This state does not have a direct prototype in the SEMAT model. Rather, it again follows the intention of the Scorecard Deck to foster preservation of practices that emerge in research work (see Table A.35). The importance of practice preservation is evident in studied project. For example, the practices for establishing inter-coder reliability developed in the Shakespeare Lives project – pilot coding, sharing examples of coding with the team, working through some coded entries with a project investigator – were reused in the InfoMigrants evaluation project. In this case it was not difficult to do without formally storing the practices since the projects went one after another with a short time gap. Yet again, since there was no formal storage of practices, it is hard to judge whether some *other* useful practices were not reused. It is also not unreasonable to suggest that should the gap in time be longer some of the reused practices could have been forgotten.

WAYS OF WORKING: INFORMED	*The principles, policies, tools and practices that inform, facilitate and shape the ways of working are selected.*
Past experience considered	The practices, policies and use cases for management tools that the team and / or its members have been exposed to in earlier comparable projects are considered.
Formalised practices considered	If the team has any form of own repository for past practices, guidance and other relevant materials, those are considered.
Institutional resources considered	If the team and / or the stakeholders have access to institutional resources on ways of working (e.g. University and / or funder policies), those are taken into account.
Relevant guidance considered	Professional and academic literature, white papers and other sources of relevant guidance on doing social data science are considered.
Team and stakeholder input taken	All the team members and, if relevant, stakeholders have been given an opportunity to give their input and have their say on the ways of working.
Team-side constraints considered	Existing constraints on the ways of working (such as principle mode of collaboration and leadership model) are considered.
Selection made	The ways of working are spelled out to a degree sufficient for the work to start.
Stakeholders: Involved	*The stakeholder representatives are contributing towards their responsibilities through engagement in the project.*
Team: Seeded	*There are initial (seed) team members and the mechanisms of expanding the team.*

Table A.31: Ways of Working: Informed. Conditions and prerequisites.

WAYS OF WORKING: EMPLOYED	*The conditions required for the work to start are met.*
Policies and practices considered	The team strives to adhere to the selected policies.
Support tools used	The teams strives to manage their activities with the selected support tools.
Discrepancies noted	The cases when the selected policies, practices, guidance and tools fail or are disruptive for the work are noted.
Gaps discovered	The areas of the work that are not covered by the existing ways of working are noted.
Benefits observed	It is clear when and under which circumstances the selected ways of working help to progress the work on the social data science project.
Work: Started	*The team has started to put its required efforts into achieving the project's goals.*

Table A.32: Ways of Working: Employed. Conditions and prerequisites.

WAYS OF WORKING: ADAPTED	*The way of working is adapted to the needs of the project and the work.*
Corrections made	The team iteratively adjusts it way of working to better match the project goals.
Improvements evident	The team can provide measurements or other form of justification that prove that the changes to the ways of working are actually beneficial.
Way of working unified	The adjusted ways of working truly apply to the work of the whole team.
Supporting resources available	All the resources of ways of working (tools, guidance) are available to every team member that might benefit from them.
Monitoring being collective	Every team member has an opportunity to participate in evaluation of the current ways of working and provide their feedback and suggestions.

Table A.33: Ways of Working: Adapted. Condition. No prerequisites specified.

WAYS OF WORKING: OPTIMISED	*The ways of working is executed fluently and efficiently with minimal costs of time and effort spent on non-productive activities.*
Practices applied effortlessly	The team does not have to put cognitive effort to work in accordance with the accepted practices.
Tools being immediately useful	The team members can instantly get the required functionality and support from the managerial tools they use.
Adaptation procedures streamlined	If the reality of the involved work dictates further changes to the ways of working, those are done quickly and efficiently.
Overheads minimised	The team can put their effort into the actual work rather than managerial overheads.
Progression rates secured	The team progresses towards the research goals and the project artefacts with a pace and quality that at least satisfy the project requirements.

Table A.34: Ways of Working: Optimised. Condition. No prerequisites specified.

WAYS OF WORKING: PRESERVED	*The new and updated practices, policies and managerial tools are preserved for further use in future social data science projects.*
Process documented	All the key aspects of the optimised ways of working - the list of the used guidances, policies, practices and tools alongside the description of how they were used - are documented.
Generalisations made	If and when possible, it is concluded when to apply and when not to apply each aspect of the ways of working in future social data science projects.
Knowledge easily available	The derived knowledge about the ways of working is stored in a manner that allows for it to be easily retrievable.
Knowledge easily navigable	The derived knowledge about the ways of working is stored in a manner that allows to conveniently navigate through its different aspects.

Table A.35: Ways of Working: Preserved. Condition. No prerequisites specified.

APPENDIX B

ETHICAL APPROVAL LETTER

The letter is provided on the next page.

University of St Andrews
Scotland's first university – 1413

University Teaching and Research Ethics Committee Sub-committee

2nd February 2016
Ilia Lvov
School of Computer Science

Ethics Reference No: *Please quote this ref on all correspondence*	CS11918
Project Title:	Towards Responsible Data Science: Participatory Design of Tools for Social Data Analysis
Researchers Name(s):	Ilia Lvov
Supervisor(s):	Dr Alexander Voss

Thank you for submitting your application which was considered at the Computer Science School Ethics Committee meeting on the 28th January 2016. The following documents were reviewed:

1. Ethical Application Form 20/11/2015
2. Participant Information Sheet 20/11/2015
3. Consent Form 20/11/2015

The University Teaching and Research Ethics Committee (UTREC) approve this study from an ethical point of view. Please note that where approval is given by a School Ethics Committee that committee is part of UTREC and is delegated to act for UTREC.

Approval is given for three years. Projects, which have not commenced within two years of original approval, must be re-submitted to your School Ethics Committee.

You must inform your School Ethics Committee when the research has been completed. If you are unable to complete your research within the 3 three year validation period, you will be required to write to your School Ethics Committee and to UTREC (where approval was given by UTREC) to request an extension or you will need to re-apply.

Any serious adverse events or significant change which occurs in connection with this study and/or which may alter its ethical consideration must be reported immediately to the School Ethics Committee, and an Ethical Amendment Form submitted where appropriate.

Approval is given on the understanding that the 'Guidelines for Ethical Research Practice' https://www.st-andrews.ac.uk/utrec/guidelines/ are adhered to.

Yours sincerely

UNIVERSITY OF ST ANDREWS SCHOOL OF COMPUTER SCIENCE

PP

Convenor of the School Ethics Committee

Ccs Supervisor
School Ethics Committee

B.1 Letters of Permission

As per request from the Open University and the British Council, I include the letters from the representatives of these organisations. The Open University hosted two of the projects used as the case studies in this thesis (see Sections 3.2 and 3.4.1). A cultural programme run by the British Council was the subject of evaluation in the first of those projects. The letters permit to submit this thesis and these case studies as its part. They can be found on the next page.

From: **Marie.Gillespie** marie.gillespie@open.ac.uk
Subject: Permissions for PhD Thesis - Ilia Lvov
Date: 10 June 2019 at 12:31
To: Ilia Lvov il23@st-andrews.ac.uk

Dear Ilia,

Thank you for the opportunity to review your thesis and for making the corrections and amendments requested by the British Council and The Open University.

I am happy to report that both parties are now satisfied that the amendments have been duly made (see email below).

We wish you all the very best of luck with the submission and viva exam.

Professor Marie Gillespie
Faculty of Arts and Social Sciences
The Open University
Walton Hall
Milton Keynes
MK7 6AA

From: "Butler, Cortina (Arts)" <Cortina.Butler@britishcouncil.org>
Date: Thursday, 23 May 2019 at 10:38
To: Marie Gillespie <marie.gillespie@open.ac.uk>
Subject: Thesis - Ilia Lvov

Dear Prof. Gillespie and Ilia Lvov

Thank you for the opportunity to review Ilia Lvov's thesis which we found very interesting.
Once the amendments requested (and outlined in a separate email) have been made, we are happy for him to proceed to submission.

Yours truly,

Cortina Butler

Cortina Butler I Director Literature I Arts
British Council I 10 Spring Gardens I London I SW1A 2BN I United Kingdom

T +44 (0)207 389 4649 I M +44 (0)7775 024 807 I BCTN (internal) 8104649
mailto:cortina.butler@britishcouncil.org

http://www.britishcouncil.org

The British Council is the United Kingdom's international organisation for cultural relations and educational opportunities. A registered charity: 209131 (England and Wales) SC037733 (Scotland). This message is for the use of the intended recipient(s) only and may contain confidential information. If you have received this message in error, please notify the sender and delete it. The British Council accepts no liability for loss or damage caused by viruses and other malware and you are advised to carry out a virus and malware check on any attachments contained in this message.

REFERENCES

Agan, T. (2007). Silent marketing: Micro-targeting. White paper, Penn, Schoen and Berland Associates.

Aitken, A. and Ilango, V. (2013). A comparative analysis of traditional software engineering and agile software development. In *2013 46th Hawaii International Conference on System Sciences*, pages 4751–4760. IEEE.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, 16(7).

Anderson, R. J. (1994). Representations and requirements: The value of ethnography in system design. *Human-Computer Interaction*, 9(2):151–182.

Apache Foundation (n.d.). Spark documentation. Available at: `https://spark.apache.org/documentation.html`. Accessed: 2015-07-04.

Ari, I., Olmezogullari, E., and Çelebi, Ö. F. (2012). Data stream analytics and mining in the cloud. In *2012 IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 857–862. IEEE.

Arthur, W. B. (1994). *Increasing returns and path dependence in the economy*. Economics, cognition, and society. University of Michigan Press, Ann Arbor.

Bacon, F. (1620). *The new organon*. Cambridge University Press, Cambridge [U.K.]. 2000 edition by Lisa Jardine and Michael Silverthorne.

Ball, A. (2012). *Review of data management lifecycle models*. REDm-MED. University of Bath, Bath, UK, 1.0 edition.

Bancroft, A. (2009). *Drugs, intoxication, and society*. Polity Press, Cambridge.

Bancroft, A. and Reid, P. S. (2016). Concepts of illicit drug quality among darknet market users: Purity, embodied experience, craft and chemical knowledge. *International Journal of Drug Policy*, 35:42–49. Drug Cryptomarkets.

Baraniuk, C. (2013). The civic hackers reshaping your government. *New Scientist*, 218(2923):36–39.

Barbu, O. (2014). Advertising, microtargeting and social media. *Procedia – Social and Behavioral Sciences*, 163:44–49.

Baron, D. P. (2001). Private politics, corporate social responsibility, and integrated strategy. *Journal of Economics & Management Strategy*, 10(1):7–45.

Barratt, M. J. (2011). Discussing illicit drugs in public internet forums: Visibility, stigma, and pseudonymity. In *Proceedings of the 5th International Conference on Communities and Technologies*, C&T 2011, pages 159–168, New York, NY, USA. ACM.

Bellamy, A. J. (2010). *Global politics and the responsibilty to protect: From words to deeds*. Routledge, London.

Bellemare, A. (2017). The secret life of Alexandre Cazes, alleged dark web mastermind. *CBC News*. Available at https://www.cbc.ca/news/canada/montreal/alexandre-cazes-millionaire-cars-property-alphabay-1.4215894. Accessed: 2019-01-29.

Bengtsson, H. (2018). future: Unified parallel and distributed processing in R for everyone. *CRAN*. Available at: https://cran.r-project.org/web/packages/future/index.html. Accessed: 2018-11-28.

Berk, R. A., Sorenson, S. B., and Barnes, G. (2016). Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *Journal of Empirical Legal Studies*, 13(1):94–115.

Bessi, A. and Ferrara, E. (2016). Social bots distort the 2016 us presidential election online discussion. *First Monday*, 21(11).

Blau, I. and Caspi, A. (2009). Sharing and collaborating with Google Docs: The influence of psychological ownership, responsibility, and student's attitudes on outcome quality. In *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, pages 3329–3335. Association for the Advancement of Computing in Education (AACE).

Bokhour, B. G. (2006). Communication in interdisciplinary team meetings: What are we talking about? *Journal of Interprofessional Care*, 20(4):349–363.

Boyce, C. and Neale, P. (2006). *Conducting in-depth interviews: A guide for designing and conducting in-depth interviews for evaluation input*. Pathfinder International, Watertown, MA.

boyd, d. (2014). What does the Facebook experiment teach us? Growing anxiety about data manipulation. *Medium*. Available at: https://medium.com/message/what-does-the-facebook-experiment-teach-us-c858c08e287f. Accessed: 2018-08-12.

boyd, d. (2018). You think you want media literacy... Do you? *Data & Society: Points*. Available at: https://points.datasociety.net/you-think-you-want-media-literacy-do-you-7cad6af18ec2. Accessed: 2018-03-26.

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

Bracken, L. J. and Oughton, E. A. (2006). 'What do you mean?' The importance of language in developing interdisciplinary research. *Transactions of the Institute of British Geographers*, 31(3):371–382.

British Broadcasting Corporation (n.d.). 5 live Hit List. Avaiable at http://www.bbc.co.uk/programmes/b04p59vr. Accessed: 2018-10-20.

British Council (2016a). Annual report and accounts 2015–16. Available at: https://www.britishcouncil.org/sites/default/files/annual-report-2015-2016.pdf. Accessed: 2017-05-10.

British Council (2016b). Corporate plan 2016–20. Available at: https://www.britishcouncil.sg/sites/default/files/corporate-plan-2016-20.pdf. Accessed: 2017-05-10.

Bruns, A. and Stieglitz, S. (2014). Twitter data: What do they represent? *it – Information Technology*, 56(5).

Buckland, M. and Gey, F. (1994). The relationship between recall and precision. *Journal of the American society for information science*, 45(1):12–19.

Burns, R. (2015). Rethinking big data in digital humanitarianism: Practices, epistemologies, and social relations. *GeoJournal*, 80(4):477–490.

Burrows, R. and Savage, M. (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data & Society*.

Button, G. (2000). The ethnographic tradition and design. *Design studies*, 21(4):319–332.

Cadwalladr, C. (2017). The great British Brexit robbery: How our democracy was hijacked. *The Guardian*. Available at: https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy. Accessed: 2018-09-25.

Cadwalladr, C. and Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*. Available at: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election. Accessed: 2018-11-27.

Cannarella, J. and Spechler, J. A. (2014). Epidemiological modeling of online social network dynamics. Technical report, arXiv.

Carroll, J. M. (1997). Human-computer interaction: psychology as a science of design. *Annual Review of Psychology*, 48(1):61–83.

Carzo Jr, R. and Yanouzas, J. N. (1969). Effects of flat and tall organization structure. *Administrative Science Quarterly*, pages 178–191.

Cassel, L. N., Posner, M., Dicheva, D., Goelman, D., Topi, H., and Dichev, C. (2017). Advancing data science for students of all majors. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*, pages 722–722. ACM.

Cate, F. H. and Mayer-Schönberger, V. (2013). Notice and consent in a world of big data. *International Data Privacy Law*, 3(2):67–73.

Chiu, C., Ip, C., and Silverman, A. (2012). Understanding social media in china. *McKinsey Quarterly*, 2(2012):78–81.

Christensen-Szalanski, J. J. and Willham, C. F. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 48(1):147–168.

Cihon, P. and Yasseri, T. (2016). A biased review of biases in Twitter studies on political collective action. *Frontiers in Physics*, 4:34.

Cohen, M. R. and Nagel, E. (2013). *An introduction to logic and scientific method*. Harcourt, Brace and Company, New York.

Coleman, S., Göb, R., Manco, G., Pievatolo, A., Tort-Martorell, X., and Reis, M. S. (2016). How can SMEs benefit from big data? Challenges and a path forward. *Quality and Reliability Engineering International*, 32(6):2151–2164.

Collins, H. (2013). Three dimensions of expertise. *Phenomenology and the cognitive sciences*, 12(2):253–273.

Costello, C. R., McNiel, D. E., and Binder, R. L. (2016). Adolescents and social media: Privacy, brain development, and the law. *Journal of the American Academy of Psychiatry and the Law*, 44:313–321.

Crawford, K. (2014). The test we can – and should – run on Facebook. *The Atlantic*. Available at: https://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/. Accessed: 2018-03-17.

Crowther, B. T. (2012). (Un)reasonable expectation of digital privacy. *Brigham Young University Law Review*, 2012(1):343–370.

Dahl, G. E., Dong Yu, Li Deng, and Acero, A. (2012). Context-dependen pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42.

Daneshgar, F., Low, G. C., and Worasinchai, L. (2013). An investigation of 'build vs. buy' decision for software acquisition by small to medium enterprises. *Information and Software Technology*, 55(10):1741–1750.

Das, A. S., Datar, M., Garg, A., and Rajaram, S. (2007). Google news personalization: Scalable online collaborative filtering. In *Proceedings of the 16th International Conference on World Wide Web*, pages 271–280. ACM.

Davenport, T. H., Barth, P., and Bean, R. (2012). How 'Big Data' is different. *MIT Sloan Management Review*, pages 22–24.

Davenport, T. H. and Patil, D. J. (2012). Data scientist: the sexiest job of the 21st century. *Harvard Business Review*, 90(5):70–76.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., and Menczer, F. (2016). Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion*

on World Wide Web, pages 273–274. International World Wide Web Conferences Steering Committee.

De Montjoye, Y.-A., Radaelli, L., Singh, V. K., et al. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539.

De Roure, D., Goble, C. A., Klyne, G., Roos, M., Hettne, K. M., Ruiz, J. E., Palma, R., Gómez-Pérez, J. M., Missier, P., and Belhajjame, K. (2011). Towards the preservation of scientific workflows. In *Proceedings of the 8th International Conference on Preservation of Digital Objects*.

De Vaus, D. (2001). *Research design in social research*. SAGE, Thousand Oaks, Calif.

Dean, J. and Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1):107.

Debreu, G. (1991). The mathematization of economic theory. *The American Economic Review*, 81(1):1–7.

Dennis, J. W., Gillespie, M., and O'Loughlin, B. (2015). Tweeting the Olympics: Towards a methodological framework for Big Data analysis of audience engagement during global media events. *Participations: Journal of Audience & Reception Studies*, 12(1):438–469.

Department for Business, Energy & Industrial Strategy (2017). Rigour, respect, responsibility: a universal ethical code for scientists. *GOV.UK*. Available at: https://www.gov.uk/government/publications/universal-ethical-code-for-scientists. Accessed: 2018-11-11.

Department for Digital, Culture, Media and Sport (2018a). Centre for data ethics and innovation consultation: Consultation outcome. *GOV.UK*. Available at: https://www.gov.uk/government/consultations/consultation-on-the-centre-for-data-ethics-and-innovation/centre-for-data-ethics-and-innovation-consultation#annex-b-key-reports-and-initiatives. Accessed: 2018-11-25.

Department for Digital, Culture, Media and Sport (2018b). Data ethics framework. *GOV.UK*. Available at: https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework. Accessed: 2018-08-12.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12):64–73.

Dickson, S. and Sigala, M. (2018). Why GDPR matters for research. *ESRC Blog*. Available at: https://blog.esrc.ac.uk/2018/05/25/why-gdpr-matters-for-research/. Accessed: 2019-02-15.

Diefenbach, T. (2008). Are case studies more than sophisticated storytelling? Methodological problems of qualitative empirical research mainly based on semi-structured interviews. *Quality & Quantity*, 43(6):875.

Diener, E. and Crandall, R. (1978). *Ethics in social and behavioral research*. University of Chicago Press, Chicago.

Dillman, D., Smyth, J., and Christian, L. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. Wiley.

Doctrow, C. (2018). Three kinds of propaganda, and what to do about them. *Data & Society: Points*. Available at: https://points.datasociety.net/you-think-you-want-media-literacy-do-you-7cad6af18ec2. Accessed: 2018-03-26.

Dourish, P. (2006). Implications for design. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 541–550. ACM.

Duggan, M. and Brenner, J. (2013). *The demographics of social media users, 2012*, volume 14. Pew Research Center's Internet & American Life Project, Washington, DC.

Duncan, M. J., Wunderlich, K., Zhao, Y., and Faulkner, G. (2018). Walk this way: Validity evidence of iPhone health application step count in laboratory and free-living conditions. *Journal of Sports Sciences*, 36(15):1695–1704.

Economic and Social Research Council (2018). ESRC research data policy. *ESRC Blog*. Available at: https://esrc.ukri.org/files/about-us/policies-and-standards/esrc-research-data-policy/. Accessed: 2019-08-19.

Edwards, A., Housley, W., Williams, M., Sloan, L., and Williams, M. (2013). Digital social research, social media and the sociological imagination: Surrogacy, augmentation and re-orientation. *International Journal of Social Research Methodology*, 16(3):245–260.

Eichenbaum, M. (1995). Some comments on the role of econometrics in economic theory. *The Economic Journal*, 105(433):1609–1621.

Erevelles, S., Fukawa, N., and Swayne, L. (2016). Big data consumer analytics and the transformation of marketing. *Journal of Business Research*, 69(2):897–904.

Ericsson, K. A. and Smith, J. (1991). *Toward a general theory of expertise: Prospects and limits*. Cambridge University Press, Cambridge ; New York.

European Parliament (2016). General data protection regulation. *Official Journal of the European Union*, L119:1–88. Accessed: 2018-08-12.

Facebook (2018a). Cracking down on platform abuse. Available at: https://newsroom.fb.com/news/2018/03/cracking-down-on-platform-abuse/. Accessed: 2018-11-10.

Facebook (2018b). Platform policy. Available at https://developers.facebook.com/policy/. Accessed: 2018-11-10.

Felten, E. (2014). Facebook's emotional manipulation study: When ethical worlds collide. *Freedom to Tinker*. Avaliable at: https://freedom-to-tinker.com/2014/06/30/facebooks-emotional-manipulation-study-when-ethical-worlds-collide/. Accessed: 2018-08-12.

Ferguson, G. A. (1971). *Statistical analysis in psychology and education*. McGraw-Hill, New York, NY, US, 3rd edition.

Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7):96–104.

Finkelstein, A. and Kramer, J. (2000). Software engineering: A roadmap. In *Proceedings of the Conference on the Future of Software Engineering - ICSE '00*, pages 3–22, Limerick, Ireland. ACM Press.

Fischer, F. (1990). *Technocracy and the politics of expertise*. Sage Publications, Newbury Park, Calif.

Fitz-Gerald, S. and Butkovic, L. (2018). The 10 biggest youtube channels right now. *Thrillist*. Available at: https://www.thrillist.com/entertainment/nation/top-youtube-channels-most-popular-youtubers. Accessed: 2018-10-29.

Franz, M. M. and Ridout, T. N. (2010). Political advertising and persuasion in the 2004 and 2008 presidential elections. *American Politics Research*, 38(2):303–329.

Freeman, R. E. (1984). *Strategic management: A stakeholder approach*. Cambridge University Press, Cambridge New York Melbourne Madrid Cape Town Singapore, 25th anniversary (2010) edition.

Friedman, M. (1968). The role of monetary policy. *The American Economic Review*, 58(1).

Fuchs, C. (2017). *Social media: A critical introduction*. SAGE Publications, Thousand Oaks, CA, 2nd edition.

Fuhrman, S. H. (1999). The new accountability. *CPRE Policy Briefs*. Available at: https://repository.upenn.edu/cpre_policybriefs/73. Accessed: 2019-02-10.

Gao, J., Zhang, T., and Xu, C. (2017). A unified personalized video recommendation via dynamic recurrent neural networks. In *Proceedings of the 2017 ACM on Multimedia Conference*, MM '17, pages 127–135, New York, NY, USA. ACM.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. MIT Press, Cambridge, MA.

Gent, I. P. (2013). The recomputation manifesto. *arXiv preprint arXiv:1304.3674*.

Genus, A. (2006). Rethinking constructive technology assessment as democratic, reflective, discourse. *Technological Forecasting and Social Change*, 73(1):13–26.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5):587–606.

Gillespie, M. (2017). Mobile methods. *Centre for Citizenship, Identities and Governance: Methods in Motion*. Available at: http://www.open.ac.uk/ccig/blogs/methods-in-motion-33-marie-gillespie-mobile-methods. Accessed: 2019-01-07.

Gillespie, M., Bell, S., Wilding, C., Webb, A., Fisher, A., Voss, A., Smith, A. W., Macfarlane, J., Martin, N., Foster, T., and Lvov, I. (2014). *Understanding the changing cultural value of the BBC World Service and the British Council*. Arts and Humanities Research Council, Swindon.

Gitelman, L., editor (2013). *"Raw data" is an oxymoron*. Infrastructures series. The MIT Press, Cambridge, Massachusetts ; London, England.

Glatard, T., Lewis, L. B., Ferreira da Silva, R., Adalat, R., Beck, N., Lepage, C., Rioux, P., Rousseau, M.-E., Sherif, T., Deelman, E., et al. (2015). Reproducibility of neuroimaging analyses across operating systems. *Frontiers in Neuroinformatics*, 9:12.

Gomm, R., Hammersley, M., and Foster, P. (2009). *Case study method*. SAGE Publications Ltd, London, United Kingdom.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., and Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks*, 38:16–27.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12.

Grabe, M. (2005). *Measurement uncertainties in science and technology*. Springer, Berlin, 2014 edition.

Graham, M., Hale, S. A., and Gaffney, D. (2014). Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*, 66(4):568–578.

Gray, J., Bounegru, L., Chambers, L., European Journalism Centre, and Open Knowledge Foundation, editors (2012). *The data journalism handbook*. O'Reilly Media, Sebastopol, CA, 1st ed edition.

Green, B. (2018). Data science as political action: Grounding data science in a politics of justice. *ArXiv preprint arXiv:1811.03435*.

Grinbaum, A. and Groves, C. (2013). What is "responsible" about Responsible Innovation? Understanding the ethical issues. In Owen, R., Bessant, J., and Heintz, M., editors, *Responsible innovation*, pages 119–142. John Wiley & Sons, Ltd, Chichester, UK.

Guillemin, M. and Gillam, L. (2004). Ethics, reflexivity, and "ethically important moments" in research. *Qualitative Inquiry*, 10(2):261–280.

Habermas, J. (1991). *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT Press, Cambridge, Massachusetts.

Halfpenny, P. and Procter, R. (2015). Introduction and overview. In Halfpenny, P. and Procter, R., editors, *Innovations in digital research methods*. Sage, Los Angeles.

Hallowell, R. (1996). The relationships of customer satisfaction, customer loyalty, and profitability: An empirical study. *International Journal of Service Industry Management*, 7(4):27–42.

Harker, S. D., Eason, K. D., and Dobson, J. E. (1993). The change and evolution of requirements as a challenge to the practice of software engineering. In *Proceedings of the IEEE International Symposium on Requirements Engineering*, pages 266–272. IEEE.

Hawkins, K. (2016). How I lost my confidence when I lost my face. *BBC News*. Available at: https://www.bbc.co.uk/news/disability-35231120. Accessed: 2018-10-29.

Henderson, J. K. (2005). Language diversity in international management teams. *International Studies of Management & Organization*, 35(1):66–82.

Hine, C. (2015). *Ethnography for the Internet: Embedded, embodied and everyday*. Bloomsbury Academic, An imprint of Bloomsbury Publishing Plc, London ; New York.

Hughes, T. P. (1993). *Networks of power: Electrification in Western society, 1880 – 1930*. John Hopkins Univ. Press, Baltimore, Md., softshell books edition.

Hutchings, S., Gillespie, M., Yablokov, I., Lvov, I., and Voss, A. (2015). Staging the Sochi winter Olympics 2014 on Russia Today and BBC World News: From soft power to geopolitical crisis. *Participations: Journal of Audience Reception Studies*.

IEEE (2017). P7003: approved Project Authorization Request. Available at: https://development.standards.ieee.org/get-file/P7003.pdf?t=92338900003. Accessed: 2018-08-12.

Inagaki, T. and Sheridan, T. B. (2012). Authority and responsibility in human–machine systems: Probability theoretic validation of machine-initiated trading of authority. *Cognition, technology & work*, 14(1):29–37.

Ivar Jacobson International (2017). Scrum Inc. is partnering with Ivar Jacobson International to build an OMG Essence Standard glossary for Scrum Terms. *Pressat.* Available at: http://tiny.cc/d60oaz. Accessed: 2019-02-18.

Jacobson, I., Meyer, B., and Soley, R. (2009). Software engineering method and theory: A vision statement. *SEMAT*. Available at: http://semat.org/documents/20181/27952/SEMAT-vision.pdf. Accessed: 2019-08-18.

Jacobson, I., Ng, P.-W., McMahon, P. E., Spence, I., and Lidman, S. (2013). *The essence of software engineering: Applying the SEMAT kernel*. Addison-Wesley, Upper Saddle River, NJ.

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., and Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94.

Janert, P. K. (2010). *Data analysis with open source tools: A hands-on guide for programmers and data scientists*. O'Reilly Media, Inc.

Janssen, M., Charalabidis, Y., and Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of Open Data and Open Government. *Information Systems Management*, 29(4):258–268.

Jarvie, I. (2011). Introduction: Philosophical problems of the social sciences: Paradigms, methodology and ontology. In *The SAGE handbook of the philosophy of social sciences*, pages 1–36. SAGE Publications Ltd, London, United Kingdom.

Jasanoff, S. (2003). (No?) accounting for expertise. *Science and Public Policy*, 30(3):157–162.

Jirotka, M. and Goguen, J. A., editors (1994). *Requirements engineering: Social and technical issues*. Computers and people series. Academic Press, London.

Kakati, S. (2017). What is data science and what is it not? *Towards Data Science*. Available at: https://towardsdatascience.com/what-is-data-science-and-what-is-it-not-c6a09d735f02. Accessed: 2018-03-26.

Keenan, T. (1997). *Fables of responsibility: Aberrations and predicaments in ethics and politics*. Meridian: Crossing aesthetics. Stanford University Press, Stanford, Calif.

Kell, D. B. and Oliver, S. G. (2004). Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays*, 26(1):99–105.

Kerzner, H. (2003). *Project management: A systems approach to planning, scheduling, and controlling*. Wiley, Hoboken, NJ, 8th edition.

Kim, G.-H., Trimi, S., and Chung, J.-H. (2014). Big-data applications in the government sector. *Communications of the ACM*, 57(3):78–85.

Kim, M., Zimmermann, T., DeLine, R., and Begel, A. (2016). The emerging role of data scientists on software development teams. In *Proceedings of the 38th International Conference on Software Engineering – ICSE '16*, pages 96–107, Austin, Texas. ACM Press.

Kirkpatrick, K. (2016). Battling algorithmic bias: How do we ensure algorithms treat us fairly? *Commun. ACM*, 59(10):16–17.

Kitchin, R. (2013). Big Data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, 3(3):262–267.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1).

Kitchin, R. and McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, 3(1):2053951716631130.

Klein, J. T. (1993). *Interdisciplinarity: History, theory, and practice*. Wayne State University Press, Detroit, 3rd print edition.

Knights, D. and O'Leary, M. (2006). Leadership, ethics and responsibility to the other. *Journal of Business Ethics*, 67(2):125–137.

Korolova, A. (2010). Privacy violations using microtargeted ads: A case study. In *2010 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 474–482. IEEE.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, pages 5802–5805.

Kozlowski, S. W. J., Gully, S. M., Salas, E., and Cannon-Bowers, J. A. (1996). Team leadership and development: Theory, principles, and guidelines for training leaders and teams. In Beyerlein, M., Johnson, D., and Beyerlein, S., editors, *Advances in interdisciplinary studies of work teams: Team leadership*, volume 3, pages 253–291. Elsevier Science/JAI Press, US.

Kramer, A. D. I., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.

Krishna, S. (2018). Facebook puts more limits on developer access to user data. *Engadget*. Available at: https://www.engadget.com/2018/07/02/facebook-limit-api-developer-app-data/. Accessed: 2019-01-08.

Krishnan, K. and Rogers, S. P. (2015). *Social data analytics: Collaboration for the enterprise*. Elsevier, Morgan Kaufmann, Amsterdam u.a.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600. ACM Press.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6:70.

Latour, B. (2009). Tarde's idea of quantification. In *The social after Gabriel Tarde: Debates and assessments*, pages 147–154. Routledge, United Kingdom.

Lazar, J. (2010). *Research methods in human-computer interaction*. Wiley, Chichester, West Sussex, U.K.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lee, S.-Y., Hansen, S. S., and Lee, J. K. (2016). What makes us click "like" on Facebook? examining psychological, technological, and motivational factors on virtual endorsement. *Computer Communications*, 73:332–341.

Leek, J. (2013). The key word in "Data Science" is not Data, it is Science. *Simply Statistics*. Available at: https://simplystatistics.org/2013/12/12/

the-key-word-in-data-science-is-not-data-it-is-science/. Accessed: 2018-03-26.

Leffingwell, D. and Widrig, D. (2000). *Managing software requirements: A unified approach*. The Addison-Wesley object technology series. Addison-Wesley, Reading, MA.

Levallois, C., Steinmetz, S., and Wouters, P. (2013). Sloppy data floods or precise social science methodologies? Dilemmas in the transition to data intensive research in sociology and economics. In Wouters, P., Beaulieu, A., Scharnhorst, A., and Wyatt, S., editors, *Virtual knowledge: Experimenting in the humanities and the social sciences*, pages 151–182. The MIT Press, Cambridge, Massachusetts.

Linden, G., Smith, B., and York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80.

Lohmann, S., Dietzold, S., Heim, P., and Heino, N. (2009). A web platform for social requirements engineering. In *Software Engineering (Workshops)*, volume 150, pages 309–315. Citeseer.

Lowrie, I. (2017). Algorithmic rationality: Epistemology and efficiency in the data sciences. *Big Data & Society*, 4(1).

Luthans, F. and Doh, J. P. (2015). *International management: Culture, strategy, and behavior*. McGraw-Hill, New York, NY, 9th edition.

MacKenzie, D. A. (2008). *An engine, not a camera: How financial models shape markets*. Inside technology. MIT, Cambridge, Mass.

MacKenzie, D. A. and Wajcman, J. (1999). Introductory essay: The social shaping of technology. In *The social shaping of technology*. Open University Press, Buckingham [Eng.] ; Philadelphia, 2nd edition.

Mao, H., Counts, S., and Bollen, J. (2011). Predicting financial markets: Comparing survey, news, twitter and search engine data. *arXiv preprint arXiv:1112.1051*.

Maréchal, G. (2010). Autoethnography. In Albert J. Mills, Gabrielle Durepos, E. W., editor, *Encyclopedia of case study research*, pages pages 44–45. SAGE Publications, Inc., Thousand Oaks.

Margetts, H. (2016). Understanding political turbulence: The data science of politics. In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, pages 2–2, New York, NY, USA. ACM.

Marshall, C. (2014). 15% of ALL youtube videos relate to gaming: Minecraft & pewdiepie FTW. *Tubular Insights: video marketing insights*. Available at: https://tubularinsights.com/15-per-cent-youtube-gaming-videos/. Accessed: 2018-10-29.

Martin, J. and Christin, N. (2016). Ethics in cryptomarket research. *International Journal of Drug Policy*, 35:84–91.

Martin, K. (2016). Understanding privacy online: Development of a social contract approach to privacy. *Journal of Business Ethics*, 137(3):551–569.

Marx, K. (1867). *Capital: A critique of political economy*. Penguin Books, London, 1981 edition.

Marz, N. and Warren, J. (2015). *Big data: Principles and best practices of scalable real-time data systems*. Manning, Shelter Island, NY.

Mayer-Schönberger, V. and Cukier, K. (2013). *Big data: A revolution that will transform how we live, work and think*. Murray, London.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., and Barton, D. (2012). Big data: The management revolution. *Harvard Business Review*, 90(10).

McInish, T. H. and Wood, R. A. (1992). An analysis of intraday patterns in bid/ask spreads for nyse stocks. *The Journal of Finance*, 47(2):753–764.

McIntyre, L. C. and Rosenberg, A. (2017). Introduction. In McIntyre, L. C. and Rosenberg, A., editors, *The Routledge companion to philosophy of social science*, Routledge philosophy companions. Routledge, Taylor & Francis Group, London ; New York.

McManus, J. (2004). A stakeholder perspective within software engineering projects. In *2004 IEEE International Engineering Management Conference*, volume 2, pages 880–884 Vol.2.

McNees, S. K. (1978). An empirical assessment of "new theories" of inflation and unemployment. In *After the Phillips Curve: Persistence of high inflation and high unemployment*. Federal Reserve Bank of Boston.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4):806.

Meredith, J. R. and Mantel, S. J. (2012). *Project management: A managerial approach*. Wiley, Hoboken, NJ, 8th edition.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., et al. (2014). Promoting transparency in social science research. *Science*, 343(6166):30–31.

Mordvintsev, A., Olah, C., and Tyka, M. (2015). Inceptionism: going deeper into neural networks. *Google Research Blog*. Available at: http://googleresearch.blogspot.ru/2015/06/inceptionism-going-deeper-into-neural.html. Accessed: 2015-06-24.

Morstatter, F., Pfeffer, J., and Liu, H. (2014). When is it biased?: Assessing the representativeness of Twitter's streaming API. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 555–556. International World Wide Web Conferences Steering Committee.

Narayanan, V., Howard, P. N., Kollanyi, B., and Elswah, M. (2017). Russian involvement and junk news during Brexit. Data Memo 2017.10, Project on Computational Propaganda, Oxford, UK.

Neuman, W. L. (2014). *Social research methods: Qualitative and quantitative approaches*. Pearson custom library. Pearson, Harlow, 7th (Pearson new international) edition.

Nye, J. S. (2004). *Soft power: The means to success in world politics*. Public Affairs, New York, 1st edition.

Oldroyd, D. R. (1986). *The arch of knowledge: An introductory study of the history of the philosophy and methodology of science*. Methuen.

Oliver, A. C. (2014). Storm or Spark: Choose your real-time weapon. *InfoWorld*. Available at: http://www.infoworld.com/article/2854894/application-development/spark-and-storm-for-real-time-computation.html. Accessed: 2015-07-04.

OMG (2015). Essence – kernel and language for software engineering methods. *Object Management Group Standard SMSC/15-12-02*. Available at: https://www.omg.org/spec/Essence/1.1/PDF. Accessed: 2016-08-15.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Papakyriakopoulos, O., Hegelich, S., Shahrezaye, M., and Serrano, J. C. M. (2018). Social media and microtargeting: Political data processing and the consequences for germany. *Big Data & Society*, 5(2).

Papamiltiadis, K. (2018). Enhanced developer app review and graph api 3.0 now live. *Facebook for developers*. Available at: https://developers.facebook.com/status/issues/205942813488872/. Accessed: 2019-01-12.

Patil, D. (2011). *Building data science teams*. O'Reilly Media.

Pennock, M. (2013). Web-archiving. Technical report, Digital Preservation Coalition.

Phillips, A. W. (1958). The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861 – 1957. *Economica*, 25(100):283–299.

Piper, A. (2018). Upcoming changes to the developer platform. *Twitter Developers*. Available at: https://twittercommunity.com/t/upcoming-changes-to-the-developer-platform/104603. Accessed: 2018-08-13.

Plesser, H. E. (2018). Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics*, 11:76.

Polit, D. F. and Beck, C. T. (2010). Generalization in quantitative and qualitative research: Myths and strategies. *International Journal of Nursing Studies*, 47(11):1451 – 1458.

Popper, K. (1959). *The logic of scientific discovery*. Routledge, London, 2005 edition.

Potthast, M., Chirigati, F., De Roure, D., Maye, R., and Stein, B. (2016). Taxonomy of actions toward reproducibility. In Freire, J., Fuhr, N., and Rauber, A., editors, *Reproducibility of data-oriented experiments in e-science (Dagstuhl seminar 16041). Dagstuhl Reports, 6, 1*, pages 135–138. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany.

Procter, R., Voss, A., and Lvov, I. (2015). Audience research and social media data: Opportunities and challenges. *Participations: Journal of Audience & Reception Studies*, 12(1).

Psylla, I., Sapiezynski, P., Mones, E., and Lehmann, S. (2017). The role of gender in social network organization. *PLOS ONE*, 12(12):1–21.

Raghavan, P. (2014). It's time to scale the science in the social sciences. *Big Data & Society*, 1(1).

Raschka, S. (2014). An introduction to parallel programming using python's multiprocessing module. Available at: https://sebastianraschka.com/Articles/2014_multiprocessing.html. Accessed: 2018-07-10.

Ratcliffe, J. H. (2015). Towards an index for harm-focused policing. *Policing: A Journal of Policy and Practice*, 9(2):164–182.

Richards, K. (2007). *Agile project management: Running PRINCE2 projects with DSDM Atern*. TSO, London.

Ritchie, J. and Ormston, R. (2014). The applications of qualitative methods to social research. In Ritchie, J. and Lewis, J., editors, *Qualitative research practice: A guide for social science students and researchers*. SAGE, London, 2nd edition.

Robinson, D. (2017). The incredible growth of python. *Stack Overflow Blog*. Available at: https://stackoverflow.blog/2017/09/06/incredible-growth-python/. Accessed: 2018-11-28.

Rogers, E. M. (1995). *Diffusion of innovations*. Free Press, New York, 4th ed edition.

Rosenberg, M., Confessore, N., and Cadwalladr, C. (2018). How Trump consultants exploited the Facebook data of millions. *The New York Times*. Available at: https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html. Accessed: 2018-11-25.

Rowley, J. (2007). The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science*, 33(2):163–180.

Ruppert, E. (2013). Rethinking empirical social sciences. *Dialogues in Human Geography*, 3(3):268–273.

Ruths, D. and Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346(6213):1063–1064.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Sahlqvist, S., Song, Y., Bull, F., Adams, E., Preston, J., and Ogilvie, D. (2011). Effect of questionnaire length, personalisation and reminder type on response rate to a complex postal survey: Randomised controlled trial. *BMC Medical Research Methodology*, 11(1):62.

Savage, M. (2010). *Identities and social change in Britain since 1940: The politics of method*. Oxford University Press, Oxford ; New York, NY.

Savage, M. and Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, 41(5):885–899.

Schedler, A. (1999). Conceptualizing accountability. In Schedler, A., editor, *The self-restraining state: Power and accountability in new democracies*. Lynne Rienner Publ, Boulder, CO.

Schot, J. and Rip, A. (1997). The past and future of constructive technology assessment. *Technological Forecasting and Social Change*, 54(2-3):251–268.

Schwaber, K. and Beedle, M. (2002). *Agile software development with Scrum*, volume 1. Prentice Hall Upper Saddle River.

Schwaber, K. and Sutherland, J. (2017). The definitive guide to scrum: The rules of the game. *Scrum Guides*. Available at: https://www.scrumguides.org/docs/scrumguide/v2017/2017-Scrum-Guide-US.pdf. Accessed: 2019-02-04.

SEMAT Inc. (n.d.). Essence user guide. View 1: Essence Lite. Available at: http://semat.org/view-1-essence-lite. Accessed: 2019-02-01.

Seo, S., Chan, H., Brantingham, P. J., Leap, J., Vayanos, P., Tambe, M., and Liu, Y. (2018). Partially generative neural networks for gang crime classification with partial information. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 257–263. ACM.

Servia-Rodríguez, S., Rachuri, K. K., Mascolo, C., Rentfrow, P. J., Lathia, N., and Sandstrom, G. M. (2017). Mobile sensing at the service of mental well-being: A large-scale longitudinal study. In *Proceedings of the 26th International Conference on World Wide Web*, pages 103–112. International World Wide Web Conferences Steering Committee.

Shvachko, K., Kuang, H., Radia, S., and Chansler, R. (2010). The Hadoop Distributed File System. In *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, pages 1–10.

SICSA (2017). DemoFest programme 2017. Available at: https://www.sicsa.ac.uk/wp-content/uploads/2014/01/DemoFest_2017_Programme.pdf. Accessed: 2018-12-11.

Siegmund, J., Siegmund, N., and Apel, S. (2015). Views on internal and external validity in empirical software engineering. In *Proceedings of the 37th International Conference on Software Engineering - Volume 1*, ICSE '15, pages 9–19, Piscataway, NJ, USA. IEEE Press.

Silverman, D., editor (2011). *Qualitative research: Issues of theory, method, and practice*. Sage, Los Angeles, 3rd edition.

Silverman, D. (2017). Getting feedback. In *Doing qualitative research*. SAGE Publications Ltd, Thousand Oaks, California, 5th edition.

Sloan, L. (2017). Who tweets in the United Kingdom? Profiling the Twitter population using the British Social Attitudes Survey 2015. *Social Media + Society*, 3(1).

Sloan, L. and Quan-Haase, A., editors (2017). *The Sage handbook of social media research methods*. SAGE Inc, Thousand Oaks, CA, 1st edition edition.

Smith, G. T. (2005). On construct validity: Issues of method and measurement. *Psychological assessment*, 17(4):396.

Smith, K. (2010). Research, policy and funding–academic treadmills and the squeeze on intellectual spaces. *The British Journal of Sociology*, 61(1):176–195.

Sommerville, I. (2008). Construction by configuration: Challenges for software engineering research and practice. In *19th Australian Conference on Software Engineering 2008*, pages 3–12.

Sommerville, I. (2016). *Software engineering*. Pearson, Boston, tenth edition, global edition.

Spiliopoulou, A.-Y., Mahony, S., Routsis, V., and Kamposiori, C. (2014). Cultural institutions in the digital age: British Museum's use of Facebook Insights. *Participations*, 11(1):286–303.

Stahl, B. C., Eden, G., and Jirotka, M. (2013). Responsible research and innovation in information and communication technology: Identifying and engaging with the ethical implications of ICTs. In Owen, R., Bessant, J., and Heintz, M., editors, *Responsible innovation*, pages 199–218. John Wiley & Sons, Ltd, Chichester, UK.

Stapleton, J. (1999). DSDM: Dynamic systems development method. In *Proceedings of 1999 Technology of Object-Oriented Languages and Systems*, pages 406–406. IEEE.

Steinmetz, G., editor (2005). *The politics of method in the human sciences: Positivism and its epistemological others*. Politics, history, and culture. Duke University Press, Durham.

Stephen, A. T., Sciandra, M., and Inman, J. (2015). Is it what you say or how you say it? How content characteristics affect consumer engagement with brands on Facebook. *SSRN Electronic Journal*. Saïd Business School WP 2015-19.

Stiglitz, J. E. (2000). Formal and informal institutions. In Partha Dasgupta, I. S., editor, *Social capital: A multifaceted perspective*, pages 59–68. The World Bank, Washington, DC.

Stilgoe, J., Lock, S. J., and Wilsdon, J. (2014). Why should we promote public engagement with science? *Public Understanding of Science*, 23(1):4–15.

Stilgoe, J., Owen, R., and Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, 42(9):1568–1580.

Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., and Lehmann, S. (2014). Measuring large-scale social networks with high resolution. *PloS One*, 9(4):e95978.

Strauss, A. and Corbin, J. M. (1990). *Basics of qualitative research: Grounded theory procedures and techniques.* Sage Publications, Inc, Thousand Oaks, CA.

Stucchio, C. (2013). Don't use Hadoop - your data isn't that big. Available at `http://www.chrisstucchio.com/blog/2013/hadoop_hatred.html`. Accessed: 2014-03-25.

Sverrisdottir, H. S., Ingason, H. T., and Jonasson, H. I. (2014). The role of the Product Owner in Scrum – comparison between theory and practices. *Procedia – Social and Behavioral Sciences*, 119:257–267.

Sysomos (2015). Software service agreement. Available at: `https://sysomos.com/software-as-service-agreement/`. Accessed: 2018-10-01.

Sysomos (n.d.). Search: Cut through the noise to make smart, data driven decisions. Available at: `https://sysomos.com/platform/search/`. Accessed: 2018-10-01.

Tansley, S., Tolle, K. M., and Hey, T. (2009). *The fourth paradigm: Data-intensive scientific discovery*. Microsoft Research.

Taras, M. (2008). Summative and formative assessment: Perceptions and realities. *Active Learning in Higher Education*, 9(2):172–192.

Tashakkori, A., Teddlie, C., and Johnson, B. (2015). Mixed methods. In Wright, J. D., editor, *International encyclopedia of the social & behavioral sciences*, pages 618 – 623. Elsevier, Oxford, 2nd edition.

Tauber, A. I. (2005). *Patient autonomy and the ethics of responsibility*. Basic bioethics. MIT Press, Cambridge, Mass.

Tauberer, J. (2014). *Open Government Data*. Joshua Tauberer, second edition.

Taylor, A. S., Lindley, S., Regan, T., and Sweeney, D. (2014). Data and life on the street. *Big Data & Society*, 1(2).

Thelwall, M. (2017). The heart and soul of the Web? Sentiment strength detection in the social web with SentiStrength. In Holyst, J. A., editor, *Cyberemotions*, pages 119–134. Springer International Publishing, Cham.

Thielman, S. (2016). Facebook fires trending team, and algorithm without humans goes crazy. *The Guardian*. Available at: https://www.theguardian.com/technology/2016/aug/29/facebook-fires-trending-topics-team-algorithm. Accessed: 2018-10-29.

Törnberg, P. and Törnberg, A. (2018). The limits of computation: A philosophical critique of contemporary Big Data research. *Big Data & Society*, 5(2).

Townsend, A. M. (2013). *Smart cities: Big data, civic hackers, and the quest for a new utopia*. WW Norton & Company, New York.

Twitter (2017). Developer policy. Available at: https://developer.twitter.com/en/developer-terms/policy.html. Accessed: 2018-11-10.

Tygart, C. (1988). Public school vandalism: Toward a synthesis of theories and transition to paradigm analysis. *Adolescence*, 23(89):187–200.

Urma, R.-G., Fusco, M., and Mycroft, A. (2015). *Java 8 in action: Lambdas, streams, and functional-style programming*. Manning, Shelter Island.

van Merkerk, R. O. and Smits, R. E. (2008). Tailoring CTA for emerging technologies. *Technological Forecasting and Social Change*, 75(3):312–333.

van Oudheusden, M. (2014). Where are the politics in responsible innovation? European governance, technology assessments, and beyond. *Journal of Responsible Innovation*, 1(1):67–86.

Vickers, J. (2014). The problem of induction. In Zalta, E. N., editor, *The Stanford encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford, CA, fall 2014 edition.

Vlaanderen, K., Jansen, S., Brinkkemper, S., and Jaspers, E. (2011). The agile requirements refinery: Applying SCRUM principles to software product management. *Information and Software Technology*, 53(1):58–70.

von Schomberg, R. (2011). Prospects for technology assessment in a framework of responsible research and innovation. In Dusseldorp, M. and Beecroft, R., editors, *Technikfolgen abschätzen lehren*, pages 39–61. VS Verlag für Sozialwissenschaften, Wiesbaden.

Voss, A., Lvov, I., and Lewis, J. (2017). The small big data manifesto. *Small Big Data*. Available at: https://smallbigdata.github.io/manifesto.html. Accessed: 20-10-2018.

Voss, A., Lvov, I., and Thomson, S. D. (2016). Data storage, curation and preservation. In Sloan, L. and Quan-Haase, A., editors, *The SAGE handbook of social media research methods*. SAGE Publications.

Voss, A., Procter, R., Slack, R., Hartswood, M., and Rouncefield, M. (2009). Design as and for collaboration: Making sense of and supporting practical action. In *Configuring user-designer relations*, pages 31–58. Springer.

Waller, N. G. (2004). The fallacy of the null hypothesis in soft psychology. *Applied and Preventive Psychology*, 11(1):83–86.

Wasserman, L. A. (2004). *All of statistics: A concise course in statistical inference*. Springer Science & Business Media.

Wetherell, M. (1998). Positioning and interpretative repertoires: Conversation analysis and post-structuralism in dialogue. *Discourse & Society*, 9(3):387–412.

Whitehead, M., Pennington, A., Orton, L., Nayak, S., Petticrew, M., Sowden, A., and White, M. (2016). How could differences in 'control over destiny' lead to socio-economic inequalities in health? A synthesis of theories and pathways in the living environment. *Health & Place*, 39:51–61.

Wholey, J. S. (1996). Formative and summative evaluation: Related issues in performance measurement. *Evaluation Practice*, 17(2):145 – 149.

Wilde, C. (2010). There is no such thing as Social Science: In defence of Peter Winch – by Phil Hutchinson, Rupert Read and Wes Sharrock. *Philosophical Investigations*, 33(2):191–199.

Williams, R. and Edge, D. (1996). The social shaping of technology. *Research Policy*, 25(6):865–899.

Willis, A., Fisher, A., and Lvov, I. (2015). Mapping networks of influence: Tracking Twitter conversations through time and space. *Participations: Journal of Audience & Reception Studies*, 12(1):494–530.

Winner, L. (1980). Do artifacts have politics? *Daedalus*, 109(1):121–136.

Wright, A. (2014). Big data meets big science. *Communications of the ACM*, 57(7):13–15.

Xu, D. (2018). API data access update. *Facebook for developers*. Available at: `https://developers.facebook.com/status/issues/205942813488872/`. First accessed: 2018-04-03. Accessed: 2019-01-10.

Yin, R. K. (2012). *Case study research and applications: Design and methods*. SAGE Publications, London.

Yin, R. K. (2013). Validity and generalization in future case study evaluations. *Evaluation*, 19(3):321–332.

Zaccaro, S. J., Rittman, A. L., and Marks, M. A. (2001). Team leadership. *The Leadership Quarterly*, 12(4):451–483.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., and Stoica, I. (2010). Spark: Cluster computing with working sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, pages 10–10, Berkeley, USA. USENIX Association.

Zhang, Y., Jin, R., and Zhou, Z.-H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52.

Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., and Lukasik, M. (2016). Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. *arXiv preprint arXiv:1609.09028*.

Zubiaga, A., Voss, A., Procter, R., Liakata, M., Wang, B., and Tsakalidis, A. (2017). Towards real-time, country-level location classification of worldwide tweets. *IEEE Transactions on Knowledge and Data Engineering*, 29(9):2053–2066.