H2020-ICT-2018-2 /ICT-28-2018-CSA

SOMA: Social Observatory for Disinformation and Social Media Analysis

From unit level to access level

Modelling academic social media data safe spaces based on administrative and genomic data management

Project Reference No	SOMA [825469]
Deliverable	D2.1: Evaluating Safe space solution including data management and processing setups
Work package	WP2: Methods and Analysis for disinformation modeling
Type	Report
Dissemination Level	Public
Date	31/08/2020
Status	Final
Authors	Lynge Asbjørn Møller, DATALAB, Aarhus University Jessica Gabriele Walter, DATALAB, Aarhus University Anja Bechmann, DATALAB, Aarhus University
Reviewers	Luca Tacchetti, Luiss Data Lab Emanuele Camarda, Luiss Data Lab
Document description	This deliverable details potential safe space solutions from other areas and evaluates the transferability to the area of disinformation monitoring in the observatory and centers.

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
v0.1	21/08/2020	Consolidation of first draft	DATALAB, Aarhus University
v0.2	28/08/2020	Review	Luiss Data Lab, LUISS University
v0.3	29/08/2020	Proofread	DATALAB, Aarhus University
v1.0	31/08/2020	Final version	DATALAB, Aarhus University

Executive Summary

Social media data offers a treasure trove for research into important research questions, such as the challenge of disinformation and its potential effects on voting behaviour and the public debate. However, privacy concerns in the wake of several data scandals have given the social media platforms an opportunity to severely restrict data access on grounds of a need to protect the privacy of the users. Thus, the need to protect privacy and researchers' need for detailed data for research are currently in conflict, and other solutions are needed that do not put legitimate researchers at the same level as malicious intruders.

This report provides an investigation of so-called research safe spaces - i.e. spaces in which authorised researchers are able to directly access and analyse potentially sensitive and identifiable data without major privacy risks for the involved subjects. The report investigates existing safe space research solutions with administrative data and genomic data, evaluates the potentials and challenges that need to be addressed in order to be inspired by such solutions to create a safe space solution for social media data, and recommends the most optimal scenarios to implement social media safe space solutions.

The investigation of existing solutions takes its starting point with Denmark as a case and explores solutions at the Danish governmental organization responsible for creating statistics on the Danish society, and the Danish governmental organization serving as a national infrastructure for genome sequencing. The results show that several aspects of the existing safe space research solutions with administrative data and genomic data are transferable to a solution for social media data, as these data types share certain characteristics when it comes to the identifiability, structure and provision of the data.

The potentials and pitfalls from the existing safe space solutions inform our recommendation to establish research safe spaces in which privacy is protected on access level rather than unit level. Previous attempts at protecting privacy on unit level by making data completely non-identifiable have caused the data to be less accurate, and we recommend that pseudonymisation should be applied instead to make it difficult to identify specific data subjects. However, focus should be on protecting privacy on access level via strict access requirements and tight rules for processing that secure access and protect the privacy and integrity of the data while facilitating further research.

The first priority scenario is the establishment of a safe space research solution facilitated outside the social media platforms, making it possible to conduct cross-platform research with potentially sensitive data without violating privacy laws, while our second priority is the establishment of safe spaces for data research facilitated by each social media platform. In both cases, it is important that the data is complete and access is administered without violating the freedom of science and hindering critical research. An important first step is testing different solutions for safe space research to secure a higher degree of balance between data utility and data confidentiality, and in the short term this would require funding, legal guidance and willingness from the platforms.

Table of Contents

List of Tables

List of Terms and Abbreviations

Abbreviation	Definition
GDPR	General Data Protection Regulation
API	Application Programming Interface

1 Introduction

In the recent decades, social media has evolved into one of the most important facilitators of information and social contact (Stieglitz et al., 2018). Thus, social media has become a major focus for researchers in multiple disciplines, and social media data potentially offers a treasure trove for research into important research questions (Batrinca & Treleaven, 2015).

However, the availability of social media data has decreased significantly in the last few years for several reasons, most notably due to privacy concerns in the wake of several data scandals such as the Cambridge Analytica scandal (Bruns, 2019). The tools offered by social media platforms for data access are subject to increasingly strict access restrictions and suffer from such major flaws that they are rendered nearly impossible to use for thorough social media research. Also, the changing levels of access causes research to be less reliable and less valid, consequently causing the policies designed and the decision taken upon this research to be equally flawed (Møller & Bechmann, 2019).

Therefore, other solutions for accessing and using social media data for research must be explored. While dedicated APIs for researchers to extract public data is certainly part of the solution to the problem, this report focuses on the more ideal so-called research safe spaces, i.e. spaces in which researchers are able to directly access and analyse sensitive and identifiable data. Such safe space solutions should not put legitimate researchers at the same level as malicious intruders, and instead mitigate the privacy risk for the involved subjects by only providing purpose-restricted access to authorised institutions and researchers with strict compliance rules.

Although such solutions still do not exist for social media data, precedence can be found in other fields than social media. For instance, genomic data contains raw DNA sequence data which is sensitive and cannot be fully anonymized, while administrative data collected by public sector organizations for record keeping on demographics, national workforce, educational level, etc. contain identifiable microdata on individual and company level. When providing research access for such data, focus is often on protecting access rather than on making the data non-identifiable, making them interesting to explore in this context.

1.1 Purpose and Scope

In this deliverable, we explore such solutions from genomic data registers and administrative data registers to evaluate the potentials and challenges that need to be addressed in order to create a similar solution for social media platform data. The investigation takes its starting point with Denmark as a case - due to the country's long history with handling of personal data - and explores solutions from *Statistics Denmark*, the Danish governmental organization responsible for creating statistics on the Danish society, and the *Danish National Genome Center*, a governmental organization serving as a national infrastructure for genome sequencing.

The investigation explores the following research questions:

- How is administrative data and genomic data respectively stored and processed in Denmark, and how is data access handled?
- How can solutions from administrative data and genomic data be transferred to social media platform data?
- What are the optimal solutions for social media platform data and which actions are needed in order to establish such solutions?

1.2 Structure of the report

In the first section of the report, we outline the reasoning behind the choice of Denmark as a case country and the choices of solutions within Statistics Denmark and the Danish National Genome Center to investigate. We also outline the methodology behind the collection of empirical data for the investigation.

In the second part of the report, we describe the safe space solutions within Statistics Denmark and the Danish National Genome Center in order to evaluate their potentials and issues of transferability if applied to a solution with social media data.

In the third and final part of the report, we outline our recommendations for the establishment of safe space social media research solutions - focusing on two different scenarios in which data is stored either outside or by the social media companies - and recommend next steps and actions needed in order to establish such solutions.

2 Methodology

In the following section, we outline the methodology behind the choices made in our investigation regarding both the selection of cases and methods to explore selected cases.

2.1 Case selection

We have chosen to explore existing solutions for safe space construction within domains outside of internet data: administrative data and genomic data. Administrative data are defined as “information collected for the purposes of registration, transaction and record keeping", typically by public sector agencies (Connelly et al., 2016, p. 3). This can include e.g. demographic, financial, educational, health, and workforce information and information from national censuses. These types of data often contain personal and sensitive information which necessitates safe research data access solutions interesting to explore in this investigation.

Genomic data refers “to raw DNA sequence data (encompassing genomic sequences of individual humans, micro-organisms residing within the human body, and other organisms), to physiological data (e.g., data relating to the association between particular genetic markers and disease risk), and to phenotypic data (including elements such as de-identified subject age, ethnicity, weight, demographics, exposure, disease state, and behavioral factors) (Van Overwalle, 2014, p. 138). Albeit very different to social media data in nature, genomic data share many characteristics with social media data - in relation to e.g. difficulties with full anonymization - making genomic data interesting to explore in this context.

The investigation takes its starting point in Denmark as a case for several reasons. Generally, the Nordic countries have a long tradition of collecting systematic data on their populations which can in part be attributed to the emergence and development of the Nordic welfare state, where the data has been utilized to guide decision making and improve the health and living conditions of the population (Tupasela et al., 2020).

Unlike most other countries in the world, the Nordic countries have developed state mandated population registers to generate accurate data concerning the population. These countries also have a long-standing use of personal identity codes to identify individuals across a broad range of public and private services from healthcare and public libraries to insurance (Tupasela et al., 2020).

In Denmark, the tradition of centralized data registration of citizens dates back to the beginning of the 20th-century. Residence registries were established in 1924 and a centralized person register was introduced in 1968 in which each citizen was assigned a personal identification number at birth or immigration (Wadmann & Hoeyer, 2018). While Denmark was neither the only nor the first country to implement such a centralized person register, the extent of its usage, its significance in everyday life and the continually growing archive of information produced by the

degree of linkage enabled by this personal identification number makes Denmark a highly data intensive country (Bauer, 2014). This fact, coupled with Denmark's long history of collecting, handling and storing personal data makes the country an appropriate case study for our investigation.

With Denmark as a case country, we have chosen to explore solutions for safe space construction within *Statistics Denmark* and the *Danish National Genome Center.*

Statistics Denmark is the Danish governmental organization responsible for creating and providing statistics on the Danish society - such as employment statistics, trade balance, and demographics. Statistics Denmark makes large use of public registers to produce these statistics, and they were the first ever to carry out a totally register based population and housing census in 1981 (Lange, 2014). Thus, Statistics Denmark has an extensive database of register data containing data collected from the 70s to the present. Due to the substantial research potential in such a database, Statistics Denmark has a Division of Research Services that helps researchers get safe access to data at so-called micro-level, i.e. data at an individual or corporate level. This solution is a meaningful case to explore in this investigation, as microdata can also be sensitive data and can be misused to identify persons or companies, necessitating a safe space solution to access the data.

The Danish National Genome Center is a governmental organization serving as a national infrastructure for genome sequencing, as well as a national database for genome data (Tupasela et al., 2020). The center stores all genomic information consensually supplied by the Danish population in connection with treatment in the health care system in its supercomputer system designed to collect, contain, analyse, handle and combine extremely large and diverse sets of data. Other than allowing for more precise treatment of individual patients, the data within the supercomputer also provides health scientists with more opportunities to incorporate genomic data in long-term research. The center is an obvious case study in this investigation, as it is still one of only a few centers worldwide for large scale handling and analysis of genomic data (Ringgaard, 2018).

2.2 Methods for interviews

To explore the solutions for safe space data storage and access at the two organisations, we chose to conduct expert interviews with the Office Manager at the Division of Research Services at Statistics Denmark and the Acting Director of High Performance Computing at the Danish National Genome Center. The two experts were chosen as their positions within the organisations enable them to provide answers to all our research questions.

We conducted *semi-structured interviews* with the two informants, allowing us to manage the interview with a structured interview guide while still allowing the respondent to answer the questions freely (Kvale & Brinkmann, 2008). The interview guide was structured into four main

topics: data characteristics, data storage, access requirements, and data access (Appendix 7.1). This enabled us to make sure that all key issues were addressed during the interviews.

The interviews were recorded, and the data gathered from the interviews were documented in the form of transcripts of the most important points made in relation to our research questions. The transcripts were checked by at least a second person in order to capture all important points. The data from the expert interviews were then analyzed qualitatively with a focus on the key issues of the research questions.

3 Safe space solutions from other areas

In this section, we account for the safe space solutions at Statistics Denmark and the Danish National Genome center in order to evaluate issues of transferability of such solutions to a solution for social media data. First, we describe each solution and institution and in detail and address application procedures for data access, data access solutions and control mechanisms against data misuse. Secondly, we discuss issues of transferability of such solutions to a solution for social media data focusing on data ownership, data identifiability, data structure and data provision.

3.1 Statistics Denmark

Statistics Denmark stores large amounts of data as a basis for statistics production and publication, and a large part of this data is available free of charge through its StatBank (Thygesen et al., 2011). This data is aggregated, so individual persons or companies cannot be identified. However, in many cases researchers using register data will need access to data at micro-level, i.e. data on individual, family, household, workplace or company level.

To facilitate register-based research, Statistics Denmark offers researchers access to micro data for specific research and analysis tasks through their Division of Research Services. These registers are updated at least once a year when Statistics Denmark publishes the statistics based on the register data. The data is stored on servers at the facilities of Statistics Denmark, and access is given as secure remote access to these servers from the researcher's own computer through the Internet. Access is administered centrally by the Division of Research Services.

The data is pseudonymised, i.e. social security numbers, company numbers or other identification keys have been removed or replaced with random numbers. The data is however still sensitive, as it is extremely detailed and includes personal characteristics – such as age, educational level, etc. – that can make it possible to re-identify individuals by combining the information with the right external knowledge.

3.1.1 Application procedure

Access to the data is only provided to authorised research or analysis institutions. The main user group includes employees at publicly funded research projects, employees in public research and analysis environments (i.e. universities, research institutes, ministries, government agencies, etc.) and researchers part of non-profit foundations. In the private sector, interest groups and consultant firms can also be authorised, but they cannot access micro data on other companies unless given an exception. Only Danish research environments are granted authorisation due to the fact that Statistics Denmark cannot enforce the contract effectively abroad, but foreign researchers can be given access to micro data if they are affiliated to a Danish authorised research environment. In this case, the affiliated institution is liable for any kind of misuse or non-compliance to regulations.

Research and analysis environments are authorised on the basis of specific assessments. In order to grant an authorisation, Statistics Denmark evaluates the organization carefully - especially when it is an organization from the private sector - and takes several factors into consideration, e.g. ownership, educational standard among the staff and research experience.

When the institution is authorised, any employee can apply for access to specific data that is relevant for the research or analysis task. Access is granted according to a so-called "need to know"-principle, meaning that researchers can only get access to the data needed to fulfill their research purpose.

Due to the data minimisation principle in the GDPR – i.e. personal data should be adequate, relevant, and limited to what is necessary for the purpose – it is the responsibility of Statistics Denmark as controller to assess whether the data requested is necessary to carry out the research and to ensure that irrelevant data is not made available (European Commission, n.d.). Thus, an applicant will need to specify the purpose of the task, the register data needed, the size of the population and the time period needed.

Research Services process about 1,500 new applications per year, and it will take about a month from applying to being granted access.

3.1.2 Data access

The microdata for research use is never handed over due to the sensitivity of the data, but researchers at authorised institutions are instead offered online access to the necessary pseudonymised microdata through a powerful research server placed at Statistics Denmark. To avoid issues of insufficient computational power, the research server is separated from the production network for statistical production and only contains pseudonymised micro data for research purposes.

When access is granted for a specific research task, the Division of Research Services will prepare the research data after which the data is transferred to the research server where remote access is given via the Internet. The researcher is then able to run analysis on the research server that offers several software packages for analysis, and the researcher can only retrieve aggregated data, where no identification of persons or enterprises is possible. Thus, micro data is at no time stored on computers outside of Statistics Denmark.

If it is needed for the research or analysis task, the microdata at Statistics Denmark can be merged with other data. External data must be handed over to Statistics Denmark in non-pseudonymised form, and Research Services will then merge it with the other data and pseudonymised all the original identification keys. In some instances where the microdata has to be merged with data requiring large amounts of storage and computational power - such as genomic data - Research

Services can connect their research servers with external computers containing the external data to conduct the merging, but the microdata still remains at Statistics Denmark.

3.1.3 Control mechanisms

To be granted access, the researcher has to sign an agreement with Statistics Denmark. This agreement states that all work with the micro data must take place only on the research server at Statistics Denmark and that no attempts may be made to remove micro data or identify persons or enterprises. This is considered a very serious breach of the agreement between the researcher and Statistics Denmark and will result in temporary or permanent withdrawal of the respective institution's authorisation.

All communications via the Internet is protected and encrypted to secure against unauthorised access. All interaction with the research server is logged to be able to detect misuse, and the Division of Research Service is notified whenever data is removed from the research server.

3.2 Danish National Genome Center

The Danish National Genome Center is a government agency and authority within the Danish Healthcare system established rather recently in 2019. With the rise of new technologies and increased knowledge about genomes, the center was established as a national wide solution for storing and allowing access to genomic data. Even though not all planned activities of the Center are established yet, its technical solution for data storage and analysis of highly complex data can still serve as a model for social media data.

The Danish National Genome Center has established a centralized infrastructure that collects and stores data from all the regions within Denmark with the aim of providing data on about 60.000 Danish people within the next four years. The primary objective of this infrastructure is to enable whole-genome sequencing to improve diagnostics and targeted treatments for patients, but the combination of Denmark’s longitudinal epidemiological datasets from the health records with genomic databases also enables new research approaches and perspectives for health care.

Therefore, the Center makes the data accessible to researchers to facilitate research about how diseases progress, how they can be detected earlier, and how they can be treated better. This system allows for centralized access to the data in a closed environment.

3.2.1 Application procedure

The Center is in the process of establishing a protocol to specify who can access and how someone can access the data via the supercomputer. Genomic data is sensitive data in multiple ways, as it provides information about individual characteristics such as ethnicity and potential health issues. This sensitive information is not removed from the data, but the data is pseudonymized to make it more difficult to re-identify individuals. Methods of introducing controlled noise to the data that

completely hinders deanonymization - such as differential privacy - are not established enough to allow for valid results and high data protection simultaneously, according to the center.

Thus, there is a high need to control data access. The protocol will follow ethical and legal standards and will require for example project approvals to protect this sensitive data. The approach is a “zero-risk” approach, and high standards for data access will be implemented. However, the system also relies on the responsibility of the user to solve the paradox between data usability and data protection. The higher the sensitivity of the data - e.g. it will be possible to request to add the social security number to the data - the stricter the precautions.

3.2.2 Data access

The infrastructure of the Center is based on a data lake system. A data lake is a large scale storage repository that is designed to handle large amounts of unstructured, semi-structured and structured data that can be processed, stored and accessed in near real-time (Miloslavskaya & Tolstoy, 2016). Such data lakes are already set up in other fields for example regarding geospatial datasets (Skluzacek et al., 2016) or within enterprises (Llave, 2018). They usually have a flat architecture and should be implemented with a metadata management that e.g. takes information about structure and semantics of the data into consideration (Khine & Wang, 2018; Mehmood et al., 2019; Ravat & Zhao, 2019). Thus, a data lake supports different kinds of data and ways of data processing (Fang, Huang, 2015).

The Danish National Genome Center uses a data lake, as it is the only available solution to meet the need for near-real time provision of a large number of unstructured and structured data that comes in diverse formats and from different providers. Their system allows maximal flexibility and based the solution on a similar solution established in a cooperation between the Danish Technical University (DTU) and University of Copenhagen (L. K. Andersen et al., 2018). A main reason for the Center to rely on a data lake was to provide the clinics and researchers with near real-time data provision and allow for customized analysis and data access.

The data lake of the Center allows flexible data input and can store data in various formats without prior normalizing or running of ETL (Extract, Transform, and Load) processes. The implementation of such processes would require high efforts and thus hinder near real-time data storage. Furthermore, the data lake can store data from various providers, enabling data from different sources to be linked in one pool. This is crucial for analysis of genomic data that requires additional information from other sources in order to treat patients - e.g. results from prior examinations, living conditions, etc.

The data lake can also store large amounts of data. As each individual’s genomic data takes up to 200GB storage, the supercomputer at the Center provides 9.5 PB usable raw capacity to handle this amount of data. Genomic data is also highly unstructured data and therefore requires a solution that can handle unstructured and structured data. The data lake solution of the Genome

Center decouples analyses from the data itself in order to manage the large volumes and size and high velocity of the data.

In order to facilitate access, the infrastructure is harmonized and extensive investments have been made regarding the software stack that structures access based on a defined set of rules. Hereby, the set of rules will be adjusted to the target group for example with researchers having a different rule set than people who enter the system in order to learn or train. The Center allows both doctors and researchers to analyse the data and thus provides a wide range of software that require a spectrum from limited to advanced technical skills. Thus, the system allows for individual environments based on prior defined rules.

These environments also allow for the import of external data, and the aim is to automate the export and import system. In order to enable data linkage with external sources, markers within the original data will be available. So far, additional data from private providers have already been added to the data lake, but no additional genomic data from other sources outside Denmark or from private companies have yet been included.

The users who are granted access can examine the data lake via their own computer using the backend solution of their home institution. The Center invested in software and technology that facilitates the use and design of the interfaces and at the same time grants access in a controlled way. This access solution also fosters acceptance within the community, since large amounts of data cannot be stored on a single computer and users are used to having data access on a centralized server.

3.2.3 Control mechanisms

The Center has established several control mechanisms to secure the data. Even though logging is a widely used mechanism to control data access, it can only be implemented by the Center to a certain extent, since data arrives at a rapid speed. The Center thus relies on logging management that uses machine learning in order to identify false positives and false negatives. Abnormalities are thus automatically detected, and users are therefore warned if they are about to cross their regulations. In case that a data leak occurs, markers in the database provide information about the cloud or the user causing the leak, enabling the Center to react accordingly.

In order to execute analyses or export data, there is an automated approval mechanism. Researchers set the characteristics of their projects which will then need approval, e.g. through the project's principal investigator. The system includes a presentation interface showing what the researcher wants to do or not and then allows for declining or approving the project. Machine learning algorithms support this process by checking e.g. whether summary statistics are conducted or not or whether the file contains information the researcher has not been granted access to.

Furthermore, data is marked to allow restricted exports. The marks allow for identifying data that is brought into the system. If it can be verified that the user brought the data in, then the user is also allowed to take this data out again.

3.3 Transferability to social media data solution

In the following, potentials and issues of transferability of the above-mentioned solutions to social media data are discussed in regards to ownership, identifiability, structure, and provision. Table 1 gives an overview of the differences and similarities between genomic, administrative and social media data in these regards.

	Administrative data	Genomic data	Social media data
Data ownership	Public institution	Public institution	Private companies
Data identifiability	Highly sensitive data, difficult to anonymize	Highly sensitive data, difficult to anonymize	Highly sensitive data, difficult to anonymize
Data structure	Structured data	Unstructured data	Unstructured and structured data
Data provision	Regular intervals	Near real-time	Near real-time

Table 1: Comparison of administrative, genomic and social media data

3.3.1 Data ownership

Social media data differs greatly from administrative data and genomic data when it comes to ownership. While public institutions own both administrative data and genomic data in the two cases, social media data is owned by private social media companies, providing issues of transferability when it comes to the scope of the data and the acquisition of the data.

While administrative data and genomic data is collected in the public interest by public institutions, social media data is collected for profit by profit-oriented private social media companies. Thus, the data that is provided to researchers is often restricted in scope and constructed to benefit businesses rather than researchers. Furthermore, the scope of the data and the level of data access regularly undergoes rapid changes as new functionalities are added to the platforms or the underlying data models are changed, causing research to be less reliable and valid (Halford et al., 2018, p. 3349). Lastly, different social media platforms each have unique characteristics and therefore produce different data (Small et al., 2012), impeding the harmonization necessary for cross-platform research.

Also, issues with intellectual property rights pervade the discussion on legal dilemmas for researchers using social media data, largely because a lot of the current intellectual property regulations are outdated in the digital age and thus not easily applied (Grinvald, 2015). When it

comes to administrative data or genomic data, the respective public institutions either collect the data themselves or are able to exert some power in order to acquire data – private companies are required by law to provide requested data to Statistics Denmark – but this is not possible when harvesting social media data. Especially exposure data is difficult to retrieve, as the platforms see this data as their intellectual property due to the fact that the curated feed is produced by algorithms developed by the platforms.

In general, social media platforms have been imposing increasingly strict restrictions on data access, making it difficult - if not impossible - to acquire substantial social media data. Thus, data acquisition would rely on cooperation on the part of the social media providers, but as exemplified by the EU Code of Practice on Disinformation this may not be sufficient. Although the Code of Practice requires platforms to enable privacy-compliant access to data for fact-checking and research activities, the terms are vague and thus implemented very differently due to the co- and self-regulatory approach and the very generic requirements (European Commission, 2018). This complicates the process of acquiring data for a safe space solution outside the social media platforms which would have to rely on strict regulation requiring the platforms to hand over data instead of urging them to do so.

3.3.2 Data identifiability

Both social media data, genomic data and administrative data is highly sensitive data, and complete data anonymization of the data that totally hinders any possibility of deanonymization is however difficult to accomplish. Genomic data can reveal sensitive personal information about health and is potentially re-identifiable due to the unique features of a genome, not only for the individual but for relatives as well (Mittos et al., 2018; Wang et al., 2017), while administrative data is extremely detailed and includes personal characteristics that can make it possible to deanonymize the data, especially when combining information. This also applies to social media data due to e.g. the amount of data points available and pictures that makes it possible to disclose identities directly or indirectly.

There are different ways to address data sensitivity and safeguard data confidentiality, and in principle these methods can be divided into non-perturbative and perturbative methods (Wirth, 2016). The former diminish disclosure risk by reducing detailed information in the data, e.g. pseudonymisation that suppress specific variable values. The latter diminish disclosure risk by introducing uncertainty in the data and thereby altering the data., e.g. introducing noise to the dataset using differential privacy. Both methods lead to a loss of information, and especially perturbative methods may affect the validity of analyses (Wirth, 2016).

In the cases of Statistics Denmark and the Danish National Genome Center, non-perturbative methods are applied to protect the data. The data is pseudonymised, where all personal information is replaced with a random reference number before publication which protects privacy by making it difficult to identify specific data subjects. This pseudonymisation does however not completely eliminate any possibility for re-identification, and the Statistics Denmark

and the Danish National Genome Center mitigate this by with strict access requirements and tight rules for data processing that protect privacy and secure data access. Pseudonymisation may also be a solution for a safe space for social media data research, pseudonymising the data before inserting it to the database or pseudonymising it at the point where it is extracted from the database to be analyzed.

The Danish National Genome Center has also looked into applying perturbative methods such as differential privacy, but concluded that with the methods currently available it would not allow for valid results and high data protection simultaneously. Differential privacy is designed to make it impossible to reverse engineer the data and disclose identities by inserting bits of noise in the data (Goroff et al., 2018), but previous attempts with social media data have shown that making data completely non-identifiable while still allowing for valid results is a difficult task (Dwork, 2008). Facebook's Social Science One initiative struggled at length to try to anonymize the data on unit level through differential privacy, which caused major delays and criticism. In August, 2019, funders of the initiative and connected researchers threatened to wind down the project due to delays (Silverman, 2019), while the European Advisory Committee of Social Science One in December the same year heavily criticized the initiative and Facebook for the continuous delays (The European Advisory Committee Social Science One, 2019).

The Social Science One initiative also illustrated that attempting to eliminate every possibility to de-anonymize the data will cause the data to be less accurate and possibly even useless for research (Lakshmanan, 2019). The first version of the codebook for the URL dataset released by the Social Science One suggested access to URLs shared by 20 people or above, but audits revealed that even this cluster size was too small to eliminate the possibility of disclosing identities. Also, differential privacy may lead to a disequilibrium between countries and regions since disclosing identities within smaller countries or regions is easier than within larger countries or regions based on the same amount of data points (e.g. demographics, interests, region). Thus, an American dataset can for instance contain more data points than a Danish or Belgian one. Insisting on protecting privacy on data level is also difficult to combine with the use of graph data, the data that have historically been used the most by social media researchers. This type of data is highly important to map disinformation flows, as connections between people reveal circulation of information and the power of bridging hubs and influencers in the network.

3.3.3 Data structure

Social media data shares characteristics with both administrative data and genomic data when it comes to data structure. Social media data comes in many different data formats and includes both unstructured data, e.g. textual content, images, videos, etc., and structured data, e.g. friend/follower relationships, geolocations, etc. (Stieglitz et al., 2018). While genomic data is highly unstructured, administrative data provided by the statistical offices is, in contrast, structured data only. Handling structured data takes less effort - e.g. regarding software, computational capacity or needed skills - than handling unstructured data and hence the solution for data storage and provision regarding administrative data is less complex than for genomic data.

Furthermore, data size of genomic data extends the one of administrative data by far, which is also the case for social media data that implies large datasets and a need for high storage capacity. The complexity of the data thus suggests a solution for safe space construction closer to genomic than to administrative data.

3.3.4 Data provision

The data provision solution at the Danish National Genome Center provides clinics and researchers with near real-time genomic data, while the administrative data at Statistics Denmark is updated in larger intervals with a time lag between data collection and data publishing. In this regard, the requirements to perform social media data research share more resemplies with that of genomic research. Social media data is produced with high velocity and in very large amounts, and a safe space data research solution would thus have to be updated with new data in regular intervals to allow for time sensitive research.

Both the Danish National Genome Center and Statistics Denmark allow researchers to access data via a remote access, but the infrastructure of the Danish National Genome Center is based on a data lake system to allow for near real-time data provision which is not necessary for administrative data. A data lake system is a data repository that enables the user to capture and store diverse types of raw data and make them available to access immediately (Miloslavskaya & Tolstoy, 2016). Such a technical solution would be ideal regarding social media as well, as it would allow the database to be continually updated with new data in different formats and from different platforms, thus permitting cross-platform research (Hall et al., 2018). However, social media companies have to overcome their reluctance to share data in order to realize such a scenario.

4 Safe space solutions for social media data

In this section, we describe and prioritize the most optimal scenarios in implementing safe space research solutions for social media data, and we evaluate next steps and actions needed in order to carry out the use of such solutions. We see two possible scenarios: establishing a safe space research solution outside the social media companies, or establishing safe space research solutions that are facilitated by the different social media companies. The potentials and pitfalls of these scenarios are accounted for in sections 4.2 and 4.3, but firstly we account for several challenges and recommendations relevant to both scenarios.

4.1 Requirements for both scenarios

In both scenarios, all appropriate steps should be taken to preserve social media users' privacy. However, as previously described perturbative methods such as differential privacy may affect the validity of analyses and previous attempts have shown that making data completely non-identifiable while still allowing for valid results is a difficult task (Dwork, 2008). These issues stemming from protecting privacy on unit level suggest that anonymity is a high bar to meet for a safe space research solution.

Instead, Statistics Denmark and Danish National Genome Center pseudonymise their data, and a similar solution is relevant for social media data, where names and other information can be replaced by random reference numbers. Article 89 of the GDPR, which addresses safeguards and derogations related to processing data for statistical purposes, explicitly lists pseudonymisation as a possible safeguard “for the right and freedoms of data subjects” (Cummings & Desai, 2018; European Parliament and Council of European Union, 2016).

While pseudonymised data is certainly more privacy protected than raw data, it is potentially re-identifiable with the right external knowledge or access to the right type of data (e.g. location data, unstructured text data and pictures). This should be mitigated by focusing on strict obligations and tight rules for processing that secure access and protect the privacy and integrity of the data while facilitating further research.

First of all, strict validation processes should be put in place that verify identity, affiliation, permissions and ethical clearance. It is of utmost importance that access is administered without violating freedom of science. The requirements for access can be set as high as needed to adhere to the research ethics and protect the privacy of the data subjects, but if these requirements are met access cannot be denied or delayed on other grounds such as the critical character of the research.

We recommend a two-step application procedure that mimics the application procedure in place to get access to the microdata at Statistics Denmark. As a first step, each research institution interested in access should be authorised on the bases of ownership, ethical clearances, and social

media research experience. When the institution is authorised, researchers are able to apply for access to specific data for a specific project. To respect the data minimisation principle in the GDPR (European Commission, n.d.), it should be ensured that only data that is relevant for the project is made accessible. Research access of research groups outside of authorized institutions would only be possible by affiliating the research group to an authorized institution which then is accountable for the proposed research project.

Secondly, similar tight rules for processing, storing and disseminating findings should be put in place to make sure that no attempts to re-identify the data are made and no potentially identifiable data are stored on external servers. Logs should be installed and audits carried out to make sure that these requirements are met, and any breaches should result in temporary or permanent withdrawal of the respective institution's authorisation. Also, thorough guidelines should be made available on how to safely process the data, and researchers should be informed about risk assessments and mitigation.

4.2 Scenario 1: Safe space solution outside platforms

The ideal and first priority scenario from a purely scientific point of view is the establishment of access to social media data in controlled and safe spaces facilitated outside the social media platforms. In this scenario, the data is removed or copied from the social media providers and stored and managed by a third-party.

This solution should cover data from several different social media platforms, which entails shared standards for data exchange among the platforms. Such an approach would thus have major research benefits in permitting cross-platform research using data from different social media platforms, a research discipline that is very complicated in the current social media environment (Hall et al., 2018).

In an external safe space solution, privacy and consent is not handled by the platforms, and a consent management solution must thus be established. In this regard, we can also find precedence in genomic data handling, as DNA data subjects consent on behalf of relatives to have the profile stored, similar to the fact that social media users consent on behalf of the ones they are connected to and communicate with. To comply with the GDPR, the external safe space solution will however also have to establish the possibility for data subjects to withdraw their consent - a function that the Danish National Genome Center also implemented in their data lake solution.

In terms of data provision, the optimal setup would include near real-time updates with new data from the social media platforms, as it would better allow for time sensitive research - potentially using a data lake system as the data repository similar to the safe space solution implemented at the Danish National Genome Center. Such a solution may however be unlikely due to the social media platform’s reluctance to share data, and a more realistic scenario would involve regular data dumps where the social media platforms provide new data in regular intervals.

The realisation of this scenario requires several actions to be taken:

1. Funding for storage and interface with the social media platforms should be secured.
2. Potential hosting institutions should be evaluated to find an institution capable and willing to host the data and administer access.
3. Legal foundation to protect privacy on access level rather than unit level should be secured with legal guidance from the Commission.
4. A two-step application procedure for data access should be formulated and implemented.
5. A new model for collecting data across private companies including an agreement with the social media companies to share data should be formulated and implemented. This should include a code of conduct explicitly for handling of exposure data due to its special status regarding intellectual property rights.

4.3 Scenario 2: Safe space solutions facilitated by the platforms

The second priority is the establishment of safe spaces for data research facilitated by each social media platform. These solutions should enable researchers to directly access and analyse sensitive and personally identifiable data that is not available through the public APIs.

Several platforms have established comparable solutions in the ad libraries provided by Google, Facebook and Twitter in 2018 to address mounting concerns over a lack of transparency and accountability in online political advertising and issues caused by new restrictions to API access (Leerssen et al., 2019). These present implementations however leave much to be desired. They are for the most part insufficient and limited in documentation, scope and information richness and lack an understanding of the work processes in research (see Møller & Bechmann, 2019), and the safe spaces for research should offer a much wider range of data.

To control that platforms provide complete data, it is essential to establish independent verification of the platform data by third-party auditors (The European Advisory Committee Social Science One, 2019). Opposed to the first priority scenario, the platforms still shape the informational structure and available data in this scenario which challenges the credibility of the research, as the researcher cannot know if the data provided are full datasets or if potentially harming data have been deleted by the platforms (Halford et al., 2018).

In this scenario, it is also essential that access is not administered by the platforms themselves but by independent third-party research councils with no ties to the platforms. In this regard, Facebook's comparable Social Science One initiative stands as a flawed attempt to achieve this, as the initiative left little room for academic independence, provided narrow terms of the work - elections and democracy - and included an inherently intransparent review process with veto power for Social Science One chairs with a privileged relationship with Facebook (Bruns, 2019). It is essential that access is administered without violating freedom of science and hindering critical research.

The realisation of this scenario requires several actions to be taken:

1. Legal foundation to protect privacy on access level rather than unit level should be secured with legal guidance from the Commission.
2. A new model for academic partnerships should be implemented that protects privacy on access level and does not put legitimate researchers at the same level as malicious intruders.
3. Documentation should be more exhaustive and data verification by independent auditors should be established to generate confidence in the data and research findings.
4. A two-step access procedure for data access should be implemented and administered by independent third-party research councils without violating freedom of science and hindering critical research, and liable research institutions should be provided access if requirements are met.

5 Conclusion

The results of the investigation shows that several aspects of safe space research solutions containing administrative data and genomic data respectively are relevant for the establishment of a safe space social media research solution.

This is especially the case when it comes to the identifiability of the data, as both social media data, administrative microdata and genomic data include sensitive information and are difficult to fully anonymise without affecting the validity of the results. Perturbative methods to anonymise the data such as differential privacy are thus not applied to protect privacy in the cases of administrative data and genomic data, as such methods make it difficult to retrieve valid results. Instead, pseudonymisation is applied, which entails replacing personal information with random reference numbers before publication, protecting privacy by making it difficult to re-identify specific data subjects. However, pseudonymisation does not completely eliminate any possibility for re-identification, which in both cases is mitigated by strict access requirements and tight rules for data processing that protect privacy and secure data access.

Furthermore, genomic data share certain characteristics with social media data in terms of data structure, size and provision that require high storage capacity and computational capacity. The safe space research solution for genomic data is therefore based on a data lake system that can capture, handle and store large amounts of diverse data and make them available to access immediately. A similar technical solution would be ideal for social media data to allow for continuous and imitate updates with new data in different formats and from different platforms, thus permitting cross-platform research and time sensitive research.

Based on the investigation of existing research safe spaces, we propose establishing safe spaces in which researchers can directly access and analyse potentially sensitive and identifiable data and in which privacy is protected on access level rather than unit level. Pseudonymisation is applied to make it difficult to identify specific data subjects, but focus should be on strict access requirements and validation processes and tight rules for processing that secure access and protect the privacy and integrity of the data while facilitating further research.

We propose two different scenarios to implement such solutions. The first priority scenario is the establishment of a safe space research solution facilitated outside the social media platforms, making it possible to conduct cross-platform research with potentially sensitive data without violating privacy laws. The realisation of such a solution is dependent on finding an appropriate third-party institution to store data and administer access, securing funding for storage, and formulating an agreement with the social media platforms to hand over data in pseudonymised form instead of anonymised form.

The second priority is the establishment of safe spaces for data research facilitated by each social media platform to enable researchers to access and analyse potentially sensitive data. In such a

scenario, it is essential that the data is continuously verified by independent auditors to control that platforms provide complete data and that access is administered by independent third-party research councils without violating freedom of science and hindering critical research. This scenario is also dependent on a new model with the social media platforms for academic partnerships that protects privacy on access level rather than unit level.

In both scenarios, an important first step is testing different solutions for safe space research to secure a higher degree of balance between data utility and data confidentiality. In the short term this would require funding, legal guidance and willingness from the platforms. The current SOMA Centers of Excellence - EU REMID and ALETHEIA - can be used as a starting point for sandbox testing of different solutions.

6 References

Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: A survey of techniques, tools and platforms. *AI & SOCIETY*, *30*(1), 89–116. https://doi.org/10.1007/s00146-014-0549-4

Bauer, S. (2014). From Administrative Infrastructure to Biomedical Resource: Danish Population Registries, the "Scandinavian Laboratory," and the "Epidemiologist's Dream". *Science in Context*, *27*(2), 187–213. https://doi.org/10.1017/S0269889714000040

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, *59*, 1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015

Cummings, R., & Desai, D. (2018). The Role of Differential Privacy in GDPR Compliance. *FATREC '18: Proceedings of the Conference on Fairness, Accountability, and Transparency.* FATREC '18. https://piret.gitlab.io/fatrec2018/program/fatrec2018-cummings.pdf

Dwork, C. (2008). Differential Privacy: A Survey of Results. In M. Agrawal, D. Du, Z. Duan, & A. Li (Eds.), *Theory and Applications of Models of Computation* (pp. 1–19). Springer. https://doi.org/10.1007/978-3-540-79228-4_1

European Commission. (n.d.). *How much data can be collected?* [Text]. European Commission - European Commission. Retrieved 24 June 2020, from https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr/how-much-data-can-be-collected_en

European Commission. (2018). *EU Code of Practice on Disinformation*. https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

European Parliament and Council of European Union. (2016). *Regulation (EU) 2016/679*. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

Fang, Huang. (2015). Managing Data Lakes in Big Data Era. What's a data lake and why has it became popular in data manangement ecosystem. In *2015 IEEE International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER 2015): Shenyang, China, 8—12 June 2015* (pp. 820–824). IEEE.

Goroff, D., Polonetsky, J., & Tene, O. (2018). Privacy Protective Research: Facilitating Ethically Responsible Access to Administrative Data. *The ANNALS of the American Academy of Political and Social Science*, *675*(1), 46–66. https://doi.org/10.1177/0002716217742605

Grinvald, L. (2015). Social Media, Sharing and Intellectual Property Law. *DePaul Law Review*, *64*(4). https://via.library.depaul.edu/law-review/vol64/iss4/4

Halford, S., Weal, M., Tinati, R., Carr, L., & Pope, C. (2018). Understanding the production and circulation of social media data: Towards methodological principles and praxis. *New Media*

& Society, *20*(9), 3341–3358. https://doi.org/10.1177/1461444817748953

Hall, M., Mazarakis, A., Chorley, M., & Caton, S. (2018). Editorial of the Special Issue on Following User Pathways: Key Contributions and Future Directions in Cross-Platform Social Media Research. *International Journal of Human–Computer Interaction*, *34*(10), 895–912. https://doi.org/10.1080/10447318.2018.1471575

Khine, P. P., & Wang, Z. S. (2018). Data lake: A new ideology in big data era. *ITM Web of Conferences*, *17*, 03025. https://doi.org/10.1051/itmconf/20181703025

Kvale, S., & Brinkmann, S. (2008). *InterViews: Learning the Craft of Qualitative Research Interviewing* (2nd edition). SAGE Publications, Inc.

L. K. Andersen, B. V. Thage, A. Syed, B. Andersen, P. L⃠ngreen, K. G. Nielsen, & S. Pedersen. (2018). National supercomputing in Denmark. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0179–0188. https://doi.org/10.23919/MIPRO.2018.8400035

Lakshmanan, R. (2019, October 1). *Facebook election interference study delayed over privacy concerns*. https://thenextweb.com/privacy/2019/10/01/facebook-election-interference-study-delayed-over-privacy-concerns/

Lange, A. (2014). The population and housing census in a register based statistical system. *Statistical Journal of the IAOS*, *30*(1), 41–45. https://doi.org/10.3233/SJI-140798

Leerssen, P., Ausloos, J., Zarouali, B., Helberger, N., & Vreese, C. H. de. (2019). Platform ad archives: Promises and pitfalls. *Internet Policy Review*, *8*(4). https://policyreview.info/articles/analysis/platform-ad-archives-promises-and-pitfalls

Llave, M. R. (2018). Data lakes in business intelligence: Reporting from the trenches. *Procedia Computer Science*, *138*, 516–524. https://doi.org/10.1016/j.procs.2018.10.071

Mehmood, H., Gilman, E., Cortes, M., Kostakos, P., Byrne, A., Valta, K., Tekes, S., & Riekki, J. (2019). Implementing Big Data Lake for Heterogeneous Data Sources. *2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW)*, 37–44. https://doi.org/10.1109/ICDEW.2019.00-37

Miloslavskaya, N., & Tolstoy, A. (2016). Big Data, Fast Data and Data Lake Concepts. *Procedia Computer Science*, *88*, 300–305. https://doi.org/10.1016/j.procs.2016.07.439

Mittos, A., Malin, B., & De Cristofaro, E. (2018). Systematizing Genome Privacy Research: A Privacy-Enhancing Technologies Perspective. *ArXiv:1712.02193 [Cs]*. http://arxiv.org/abs/1712.02193

Møller, L. A., & Bechmann, A. (2019). *D2.2: Research Data exchange (and transparency) solution with platforms*. The European Commission.

Ravat, F., & Zhao, Y. (2019). Data Lakes: Trends and Perspectives. In S. Hartmann, J. Küng, S. Chakravarthy, G. Anderst-Kotsis, A. M. Tjoa, & I. Khalil (Eds.), *Database and Expert Systems Applications* (Vol. 11706, pp. 304–313). Springer International Publishing.

https://doi.org/10.1007/978-3-030-27615-7_23

Ringgaard, A. (2018, March 2). *Forskere: Derfor skal du donere dit DNA til Nationalt Genom Center*. https://videnskab.dk/teknologi-innovation/forskere-derfor-skal-du-donere-dit-dna-til-nationalt-genom-center

Silverman, G. (2019, August 27). *Exclusive: Funders Have Given Facebook A Deadline To Share Data With Researchers Or They're Pulling Out*. https://www.buzzfeednews.com/article/craigsilverman/funders-are-ready-to-pull-out-of-facebooks-academic-data

Skluzacek, T. J., Chard, K., & Foster, I. (2016). Klimatic: A Virtual Data Lake for Harvesting and Distribution of Geospatial Data. *2016 1st Joint International Workshop on Parallel Data Storage and Data Intensive Scalable Computing Systems (PDSW-DISCS)*, 31–36. https://doi.org/10.1109/PDSW-DISCS.2016.010

Small, H., Kasianovitz, K., Blanford, R., & Celaya, I. (2012). What Your Tweets Tell Us About You: Identity, Ownership and Privacy of Twitter Data. *International Journal of Digital Curation*, *7*(1), 174–197. https://doi.org/10.2218/ijdc.v7i1.224

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

The European Advisory Committee Social Science One. (2019, December 11). *Public statement from the Co-Chairs and European Advisory Committee of Social Science One*. https://socialscience.one/blog/public-statement-european-advisory-committee-social-science-one

Thygesen, L. C., Daasnes, C., Thaulow, I., & Brønnum-Hansen, H. (2011). Introduction to Danish (nationwide) registers on health and social issues: Structure, access, legislation, and archiving. *Scandinavian Journal of Public Health*, *39*(7_suppl), 12–16. https://doi.org/10.1177/1403494811399956

Tupasela, A., Snell, K., & Tarkkala, H. (2020). The Nordic data imaginary. *Big Data & Society*, *7*(1), 2053951720907107. https://doi.org/10.1177/2053951720907107

Van Overwalle, G. (2014). Governing Genomic Data: In B. M. Frischmann, M. J. Madison, & K. J. Strandburg (Eds.), *Governing Knowledge Commons* (pp. 137–154). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199972036.003.0005

Wadmann, S., & Hoeyer, K. (2018). Dangers of the digital fit: Rethinking seamlessness and social sustainability in data-intensive healthcare: *Big Data & Society*. https://doi.org/10.1177/2053951717752964

Wang, S., Jiang, X., Tang, H., Wang, X., Bu, D., Carey, K., Dyke, S. O., Fox, D., Jiang, C., Lauter, K., Malin, B., Sofia, H., Telenti, A., Wang, L., Wang, W., & Ohno-Machado, L. (2017). A community effort to protect genomic data sharing, collaboration and outsourcing. *NPJ Genomic Medicine*, *2*. https://doi.org/10.1038/s41525-017-0036-1

Wirth, H. (2016). Analytical Potential Versus Data Confidentiality – Finding the Optimal Balance. In C. Wolf, D. Joye, T. Smith, & Y. Fu, *The SAGE Handbook of Survey Methodology* (pp. 488–501). SAGE Publications Ltd. https://doi.org/10.4135/9781473957893.n32

7 Appendix

7.1 Interview guide

Data characteristics

What are the main purposes of data collection and storage in your institution? (Aim: Information about data provision)

Do your purposes differ from those in research environments? (Aim: Information about institutional peculiarities and deviations from research)

What characterizes this type of data and how does the type of data affect the handling of the data? (Aim: Information about data characteristics)

Data storage

How is the data stored in your institution when it comes to anonymity?

How long does it take from data collection to data provision, e.g. which is the newest data available which time period do you cover within the data? (Aim: Information about scope of data content and institutional conditions)

Access requirements

What are the basic requirements to be eligible for data access, and how is the application procedure? (Aim: Information about conditions of data access)

How many applications for access do you process?

What kind of research groups have access to the data you provide? (Aim: Information about data access clients/ users)

Do you allow access to research groups outside of universities e.g. the media, business industry or private institutions, and does it follow different regulations? (Aim: Information about scope of data access)

Data access

Who is responsible for research access to the data? (Aim: Information about institutional setting)

How is the data made available to researchers? Can it be downloaded? (Aim: Information about data access management)

Can researchers merge your data with other types of data?

Can researchers get access to sensitive data without anonymization or only to anonymized data? (Aim: Information about anonymization procedures)

How do you anonymize the data and how do you control the anonymization? (Aim: Information about anonymization controls)

Does the procedures of data access vary by the sensitivity of the data in the dataset? (Aim: Information about sensitivity of data)

How flexible is the data access solution in adjusting research methods, running off-site analyses, and storing results? (Aim: Information about researchers autonomy)

Which control mechanisms exist against misuse of the data by the researchers? Are users logged? (Aim: Information about control mechanisms)

Experiences and expectations concerning social media data

What experience does your institution have in handling data from private companies eg. session logging or data from insurance companies, private doctors, banks etc.? If you have experience, how does data management for this data differ from normal procedures? (Aim: Information about connections to social media data)

How do you assess the desire for the institution to maintain and access social media data? If your institution does not want this commitment, which organization could you point to? (Aim: Information about potential collaboration)