


Open research data and privacy violations

Journal of Information Science
1–14
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01655515241297399
journals.sagepub.com/home/jis


Laila Dahabiyeh 

Department of Management Information Systems, School of Business, The University of Jordan, Jordan

Nashrawan Taha

Department of Library and Information Science, School of Educational Science, The University of Jordan, Jordan

Abstract

Open research data refer to the practice of sharing research data with others to enhance science innovations and discoveries. Despite the great potentials of open research data, it comes with certain limitations, especially with regards to the privacy of research participants' data. In this article, we examine the tension between public data repository policy and the General Data Protection Regulation (GDPR). To achieve that, we draw on privacy as contextual integrity theory. We further enrich our research by interviewing 12 researchers from European institutions to examine their perception of whether open research data have privacy challenges or not. Our findings reveal that according to the heuristics steps of privacy as contextual integrity and the GDPR requirements, open research data may entail a violation to research participants' informational privacy. Moreover, data repository's policy is geared towards protecting the confidentiality of research participants' data rather than their privacy. We further reveal that researchers conflate privacy and anonymity and lack knowledge of sharing research data practices.

Keywords

Contextual integrity; GDPR; open research data; open science; privacy

1. Introduction

The introduction of the fourth industrial revolution introduced in the 21st century has led to technological and data revolution. We are witnessing the creation and processing of vast amount of data with all types, which is changing the world of research and academia. The scientific community has now an unprecedented opportunity to benefit from the 'openness' of the new open science paradigm that is changing the culture of research. The open science movement, and more particularly the open research data (ORD), is one aspect of this technological and data shift, where sharing of research data has become a trend that hopes to help in enhancing science innovations and discoveries [1].

ORD refers to the practice of sharing research data with others to enable them to mine the data for new discoveries [1]. Advocates of ORD emphasises the replicability and democratisation of scientific knowledge that ORD facilitates through enhancing the access to research resources to wider community [2]. However, sharing research data that include personal information about research subjects in public data repositories, even in anonymised form, may threaten privacy. In this article, we examine the tension between public data repository policy and the General Data Protection Regulation (GDPR). GDPR is the most comprehensive privacy legislation in Europe and the toughest privacy law in the world. This tension is examined by drawing on privacy as contextual integrity theory [3,4].

This research seeks to answer the question: does sharing research data in public data repository entail privacy violation of research subjects? And, if so, how can the privacy and related interests be respected? To answer these questions, we apply the heuristics steps of privacy as contextual integrity. We further enrich our research by conducting interviews

Corresponding author:

Laila Dahabiyeh, Department of Management Information Systems, School of Business, The University of Jordan, Queen Rania Street, Amman 11942, Jordan.

Email: laila.dahabiyeh@ju.edu.jo

with 12 researchers from European institutions to examine their perception of whether ORD has privacy challenges or not. Our findings reveal that according to the heuristics steps of privacy as contextual integrity and the GDPR requirements, ORD may entail a violation to research participants' informational privacy. In addition, data repository's policy is geared towards protecting the confidentiality of research participants' data rather than their privacy. We further reveal that researchers conflate privacy and anonymity and lack knowledge of sharing research data practices.

The rest of the article is organised as follows. Next, we review the literature on ORD in terms of benefits and challenges including privacy issues in ORD. We then discuss governance of personal information focusing on the GDPR and its privacy requirements. The next section explains our theoretical framing, namely, privacy as contextual integrity theory, followed by applying privacy as contextual integrity theory to the practice of depositing research data in data repository (specifically the Inter-university Consortium for Political and Social Research (ICPSR)). After that, we detail our qualitative study. Finally, we discuss our findings and provide important implications.

2. Theoretical foundation

2.1. Open research data

The term ORD was generated from the broader area of open data [5]. It relates to the openness of data access and reuse where researchers can use the data produced from a scientific research freely with no resections on its re-use and sharing [5–8]. The ORD term is defined by the Open Data Institute [9] as 'Information that is available for anyone to use, for any purpose, at no cost', but under a licence of attribution and share-alike. Based on the European Commissions [10], ORD stands for the use of scientific data without any restrictions, where any other researchers can freely access and use the research data available in data repositories. Sharing of research data aims mainly to make research data more 'Findable, accessible, interoperable, and reusable (FAIR)' [7,11].

The practice of open data can be useful in increasing scientific innovation and offering a better opportunity for researchers to be recognised by increasing their visibility [12,13]. It also helps in providing transparency of research, expanding collaboration and increasing the scientific discovery [7,14,15].

However, ORD involves some limitations that exceed the benefits of providing free and easy access to research. In fact, some looked into the ORD paradigm as a 'conundrum' [1]. Researchers believe that sharing of research data could not be feasible in some fields of research [1] and hence their behaviour towards data sharing varies according to their field of research [16].

Other issues can be related to the researchers themselves and to the fact that many of them are not willing to share their research data, lack the skills and expertise needed to do so [1,5,14] or require access to tools to share research data [5,14]. Researchers might also lack the motivation to share their data if they get no credit for sharing data, such as in citation [14,17].

The issues related to sharing of research data can also be attributed to the context and the difficulty in interpreting data for a particular context [1,5,15]. This is particularly correct for research in humanities where scholars may find that their research process and approaches are too individualistic when compared with the research norms built in research repositories [15]. Prior research highlighted that many instances in qualitative research might not directly lend themselves to being captured explicitly in research data; however, they are critical for understanding the context. Accordingly, sharing qualitative data can conflict with some methodological principles of qualitative research [5].

Privacy is another important issue related to ORD. The privacy of participants and the sharing of their data is still arguable since the ownership of the data seems to be blurred [14,18]. Some argue that data ownership belongs to the universities or researchers [14]. Moreover, data privacy could not be guaranteed in the case of multiple variables and individuals could be identified and traced back, which may lead to revealing participants' identities and breaching their privacy [14]. Although various anonymisation methods exist, these vary in their effectiveness in protecting privacy and strong anonymisation techniques can result in low data utility [19]. The intentional or unintentional data 'misuse' can be a concern for sharing research data [18] as well as the problem of creating 'data gaps' where data can become 'detached' from the conclusion made in the original research [8]. Legal issues, such as copyright is another point of concern in relation to sharing research data; this is particularly true when dealing with image data in particular contexts [8,15].

The research data management (RDM) has inevitably come up after the open science initiative to help in maximising the 'potential for open data' [5]. In RDM, researchers are going to be responsible for RDM, which should start with a data management plan (DMP) that includes all the steps needed in the lifecycle of data during the entire research process [5]. The DMP should include all the issues that need to be taken into consideration when sharing research data. It also should contain how researchers would address any concern related to data sharing. The DMP also explains the options that researchers will have in making their data open and describes how to access the research data to guarantee the

quality of data and their publication. In fact, some funding agencies started to require the DMP for large grants [1]. This is mainly required because the DMP helps to ensure that data are prepared to be properly used for the wider community benefit [7]. Furthermore, publishers' policies are enforcing new requirements for publishing research by asking for data availability statement and/or depositing data in specific data repository.

Despite having open science policies, institutions are still in need to develop a policy related to ORD [11]. So, in order to facilitate the ORD process, it would be highly valuable that all stockholders involved in the issue of ORD including researchers, universities, publishers, research and technological organisations, libraries, funding agencies and research repositories, to collectively work on developing the needed policies, requirements and study the benefits and concerns involved in ORD and to come up with a conclusion on a 'global level' since the open science paradigm is going to be globally shared [8].

2.2. Governance of personal information: the GDPR

The GDPR is a comprehensive legislation to protect personal information. It came into effect in May 2018 to harmonise data privacy laws across Europe. The regulation specifies rules for protecting the processing of personal information where personal information refers to any information that can identify a person directly or indirectly, and processing means any operation performed on personal information such as collection, storing, modifying and sharing (Chapter 1, Art. 4). The responsibility for ensuring that the processing of personal information complies with GDPR requirements lies within the controller who is a 'legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data'. (Chapter 1, Art. 4).

The GDPR places strong emphasis on users' consent to the processing of their information in order to consider the processing lawful (Chapter 2, Art. 6). The regulation further mandates that the consent form must be easily accessible and use clear and plain language, and data subjects have the right to withdraw their consent at any time (Chapter 2, Art. 7). The regulation thus adopts a control view on information privacy where privacy is perceived as the ability to control what information to disclose or not disclose about oneself. Accordingly, when obtaining personal information, the controller must provide the data subject with information on its (the controller) identity, the purpose of processing and any entity that will have access to the personal information (Chapter 3, Art. 13). Moreover, if the controller decides to process the personal information for purpose other than the original one, it needs to provide the data subject, and prior to processing, information on the other purpose and any other relevant information (e.g. the recipients, the duration for which the data will be stored) (Chapter 3, Art. 13).

3. Privacy as contextual integrity

Privacy as contextual integrity theory was introduced by Nissenbaum in 2004 in an attempt to capture the privacy challenges introduced with digital innovations. According to this theory, privacy is defined as an appropriate flow of information according to context-dependent norms. The theory has been used to evaluate information flow in various contexts such as Facebook privacy policy [20], big data research [21], smart home [22] and contact-tracing applications [23].

A central tenet of contextual integrity is that all information flows are governed by norms that can be manifested in laws and regulations or societal expectations. In addition, individuals engage in different spheres or contexts (e.g. education, health care, political) where each have its own norms and expectations that govern information flow [3,4]. Nissenbaum [3] identifies two types of informational norms: norms of appropriateness and norms of distribution. The violation of either of these norms, or both, constitute violation of privacy. Norms of appropriateness specify what information is appropriate to reveal or share in a particular context. For example, it is appropriate to share financial information with one's bank, but not with one's physician or friends. Norms of distribution refers to the norms that govern information flow from one entity to another. Not only the appropriateness of information matters, but also whether the distribution of information flow respects the contextual norms of information flow. Such norms can be confidentiality and consent.

Consequently, contextual integrity perceives privacy as context dependent. This indicates that while sharing information in one context might not violate privacy, sharing the same information in another context might constitute threat to privacy. Contextual integrity identifies five parameters for evaluating the information flow: the sender of information, the recipient of information, the subject of information, the type of information and the transmission principle that determines when the information flow is permitted. A change in any of these parameters might result in violation of privacy [23]. Accordingly, contextual integrity argues that a procedural view of privacy where information flow is deemed appropriate as long as it complies with certain rules (e.g. informed consent) is not sufficient to ensure privacy [24].

The theory offers a decision heuristic to evaluate whether an action or practice may violate privacy. This heuristics involves nine steps [3]: (1) describe the new practice in terms of information flow; (2) identify the prevailing context; (3) identify information subjects, senders and recipients; (4) identify transmission principles; (5) locate entrenched informational norms; (6) prima facie assessment; (7) first evaluation; (8) second evaluation; and (9) conclusion. These steps are explained and applied to the ORD next.

3. Privacy as contextual integrity: the case of ORD

In this section, we will apply the contextual integrity heuristics steps to evaluate whether the sharing of research data by depositing it in a public data repository will entail privacy violations. As we are focusing on social science research, we have selected ICPSR as the data repository that we will examine. ICPSR is the world's largest social science data archive established in 1962 with over 250,000 files of research in social and behavioural sciences.

It is worth noting that the application of the steps was done by each researcher individually to reduce bias. After that, the analysis of the case was discussed between the researchers. The discussion revealed that both researchers reached the same conclusion.

3.1. Describe the new practice in terms of information flow

The first step of the decision heuristic entails describing how the information flows in the new context. In the context of ORD, the information flows start from having research participants providing their information to researchers according to the data collection method used (e.g. questionnaire, interviews). Researchers then share these data in a data repository such as ICPSR after creating an account. The data can then be used by other researchers according to the terms of use.

3.2. Identify the prevailing context

The prevailing context is ORD in Europe where GDPR applies.

3.3. Identify information subjects, senders and recipients

The information subjects are about whom the information is, the senders are the ones from whom the information flows and recipients are the ones to whom the information flows. In our context, information subjects are the research participants. The senders are researchers that submit participants' data into data repositories. Recipients are ICPSR, as a data repository and the researchers who access the data repository and download participants' data to use them in their own research.

3.4. Identify transmission principles

Transmission principles represent conditions and constraints on the information flow in a particular context. Accordingly, they denote whether information transfer should or should not occur. Confidentiality is one salient example of a transmission principle. In our context, ICPSR has requirements to maintain data confidentiality before researchers share or deposit their data in the repository. ICPSR acknowledges that once data are publicly shared, it is difficult to ensure that research subjects' confidentiality is protected and respected by other researchers. Accordingly, the common practice is to have information that might threaten confidentiality removed or masked before publicly sharing it, while ensuring at the same time that the data are still valuable for reproduction and re-use [25].

The GDPR goes beyond confidentiality to emphasise that the sender must inform the data subjects (i.e. research participants) of its identity, the purpose of collecting and processing the information and the identity of the recipients.

3.5. Locate entrenched informational norms

Entrenched informational norms refer to the established practices for information flow that formulate expectations of whether a flow is appropriate or not. Accordingly, informational norms entails transmission principles but further include elements of context, the type of information, the subject, the sender and the recipient. Since its publishing in 2018, the GDPR is considered the main regulation in Europe for data protection and privacy. The regulation is considered a key reference point for any entity that collects, processes or transfers personal information.

3.6. *Prima facie* assessment

In this step, the new practice is evaluated against the informational norms to determine whether contextual privacy has been violated or not. The *prima facie* assessment of the ORD against the GDPR reveals that data repositories policy, specifically ICPSR, may not comply with the GDPR requirements. The GDPR clearly states that data subjects must provide their consent on the type of information collected about them, the purpose these data will be used for and the recipients of their data (Chapter 3, Art. 13). However, the GDPR provides exception to this general rule for data collected for research purposes, which enables researchers to collect and process subjects' data beyond the purpose they agreed on when the data were first collected, which may indicate that ORD respects GDPR requirements. Article 5(1-b) Chapter 2 states that:

[personal data shall be] collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation').

Moreover, it can be argued that processing data for research purposes can be considered a legitimate interest, which falls under the lawful bases for processing that does not require further consent from research subjects. Nonetheless, the waiver for further processing of research data is subject to having 'appropriate safeguards' in place, where appropriate safeguards refer to technical and organisational measurements for protecting the data (Chapter 9, Art. 89), and the exception for 'legitimate interest' is conditioned on the expectations of research participants. Recital 47 states:

At any rate the existence of a legitimate interest would need careful assessment including whether a data subject can reasonably expect at the time and in the context of the collection of the personal data that processing for that purpose may take place.

When research participants consent to participate in a research study, they often agree on sharing their data for the purpose of the research they are participating in. The language used in the consent form emphasise this as well. Accordingly, they do not expect that their data might be used for purposes other than the purpose of 'this' research or that it will be shared with 'other' parties or be publicly available in data repositories. Therefore, according to GDPR requirements, there is a violation of privacy. The ICPSR guidelines for the informed consent is very general with neither details on how the data might be used beyond this research nor the identity of the recipients of the data. Researchers are guided to state, 'We may use the information you provide for future research studies. We will not ask for your additional informed consent for these studies' [25, p. 14]. Although this guideline informs individuals of the possibility of future use of their data, it is still problematic because it implies that the same research team will be the one who will use the data, and it gives no information on the fact that the information will be shared publicly in a data repository.

Moreover, the conventional language in informed consent with regards to purpose 'Your answers will be held in strict confidentiality and will be used only for the purposes of this study' is perceived by ICPSR [26] to be appropriate; they argue, 'the purpose of a study is broadly defined in substantive terms that are appropriate for almost all forms of research conducted through secondary analysis of the data'. However, the purpose of a research conducted to identify factors behind social media continuous intention is different than a study that aims to explore how culture influences technology-related decisions. One cannot make a deduction that participants giving their consent on the first study will automatically be fine with their data used for the other purpose. In addition, the phrase '... for the purposes of *this* study' (italic added), is misleading as it gives no indication that the data might be used in other research studies.

The matter becomes more alarming when considering 'restricted-use data' option in ICPSR. Restricted-use data are data that might lose its analytical potential if user identifying information is removed [25]. ICPSR allows restricted-use dataset to be shared with approved clients under the condition that they will abide by the confidentiality rules. This means that identifiable data about research participants is being shared with third parties without their knowledge or consent.

3.7. *Evaluation I*

If the initial assessment resulted in a violation of contextual privacy, the next step is to evaluate the political factors affected by the new practice and identify possible harms. While research participants might not be directly harmed by ORD practice, they lose control over their data threatening by this their autonomy. The GDPR seeks to empower research participants by giving them control over their data, ORD practices, however, shift this power to data senders. While the

GDPR acknowledges that the ownership of the data is a right for the research participants, giving the sender the right to take the decision of sharing the data can imply that they (data senders) own participants' information. In addition, the GDPR protects individuals' rights to access, modify and delete their data, but how this is guaranteed in ORD is not clear.

3.8. Evaluation II

The second step of evaluation focuses on how the new practice impinge on the values, goals and ends of the context itself. It is apparent that there is a tension between ORD and data protection that GDPR aims to achieve. This tension threatens societal values of autonomy, privacy as well as trust in the scientific community. For example, failing to appropriately inform research participants about how their data are being used and re-used and with whom it is being shared can result in misuse of data and a trust crisis in the scientific community. The Cambridge Analytica scandal showed how scientific research can be a disguise to manipulation to serve political interests.

3.9. Conclusion

Based on the analysis above, we conclude that ORD can present a violation of contextual privacy.

4. The qualitative study

To complement and enrich our study, we decided to explore the opinions of researchers in European institutions on the privacy issues related to ORD and its alignment with the GDPR. For that, we conducted semi-structured interviews.

4.1. Data collection

Participants were social science researchers and librarians (engaged in research) who were familiar with sharing research data via public data repositories and are aware of GDPR. Twelve researchers from the United Kingdom, Germany, France, Spain, Belgium and Turkey were chosen to answer the interview questions (see Table 1). Email invitation that included the research consent form and information sheet were sent to potential participants. Interviews were conducted with those who expressed willingness to participate in the study.

Interviews were conducted online via Zoom application. All interviews were recorded and transcribed after taking the participants' approval. A number of open-ended questions were used that explored researchers' perspectives of sharing research data, their opinions of privacy concerns associated with ORD, and their perception on the alignment between ORD and the GDPR.

4.2. Data analysis

Thematic analysis was used to analyse the interviews' data [27], where interview transcripts were imported into Atlas.ti qualitative data analysis software. The first step in thematic analysis is getting familiar with the data, which was achieved through repetitive reading of the transcripts. In the second step, we did descriptive coding where we assigned words and

Table 1. Participants' details.

Participant ID	Country	Position
R1	Belgium	Professor – Faculty of Social Sciences
R2	France	Associate Professor – School of Business
R3	United Kingdom	Senior Lecturer (Associate Professor) – School of Business
R4	United Kingdom	Senior Lecturer – Management School
R5	United Kingdom	Professor – School of Business
R6	United Kingdom	Lecturer – School of Computing Sciences
R7	United Kingdom	Subject Liaison Librarian
R8	Germany	Managing Director
R9	Turkey	Library Director
R10	United Kingdom	Senior lecturer – Information School
R11	Spain	Librarian
R12	Turkey	Research Data Coordinator

Table 2. Examples on the analysis process.

Examples from the data	Descriptive codes	Themes	Categories
'There might be management systems ... it might there is software available, I do not know enterprise system, I have no clue since I do not have the experience'. R1	No experience	Lack of knowledge	Uncertainty
'And I think we, there is a missing understanding, first of all, even I've noticed it by most of UK universities about, first of all, what is open research? And also, what are the benefits of open research and the challenges?' R3	Missing understanding		
'I don't know, here, there is kind of a gray zone, that is not perfectly crystal clear'. R2	Grey zone		
'it was not, to be honest, like made mandatory'. R6	Not mandatory	Justifications	
'I didn't really think of that too much because I was just focused on my research and on getting it done, and meeting deadlines, and doing a proper piece of work'. R7	Too much work		
'all data, all the time, should not lead to identifying someone'. R4	Identifying data	Anonymity	Privacy concerns
'you need to make sure that the either the original organization that you are citing agrees to share publicly the data or you make all everything name anonymous'. R2	Anonymous		
'I think, they [regulators] still have a lot of ways to go with AI and use of AI'. R12	Use of AI	AI technology	
'I know ... are already doing data analysis, qualitative and quantitative using ChatGPT or Bard ... those models are learning about the respondents. If they have enough data the person can be known'. R6	Models		
'any good consent form fully informs research subjects about what will happen to their data'. R1	inform	Consent form	
'Researchers must ask consent from participants to reuse data in further research and to deposit it in an archive it'. (R7)	Reuse data		
'I've got to use different consent forms that comply with the university regulations, for instance, and the university regulations, it's just GDPR'. R3	University templates		
'I use template from my university which was drafted by legal advisors'. R1	Legal advisor	University regulations	
'but we as researchers, we need to operate under the rules, right, and the policies of the university'. R4	University operation requirements		

UK: United Kingdom; AI: artificial intelligence; GDPR: General Data Protection Regulation.

short phrases to describe the data. In the third step, we grouped the descriptive codes into relevant 'themes' that describe participants' view of ORD practices and any privacy issues related to using ORD in data repositories. The themes were identified based on conceptual similarity [28], which were then grouped into two main categories, namely, uncertainty and privacy concerns, forming the main categories of our findings. The researchers read all the interview transcripts for accuracy and the transcripts were divided between the two researches for coding. Several rounds of discussion took place to avoid any bias in analysing the data and reach consensus about the findings.

It is worth noting that while the aim of the analysis was to better understand researchers' perception of the privacy concerns related to ORD and the tension (if any) between ORD and GDPR, we found that their responses to the above were highly related to their knowledge (or lack) about ORD practices. Therefore, and to maintain clarity and good flow, we also present the analysis and findings related to ORD practices.

Table 2 displays examples of the data analysis process.

5. Findings from the qualitative study

The interviews conducted with researchers from Europe revealed the uncertainty and lack of awareness about public data repositories and their data management practices as well as GDPR and its implications for research. Our findings are presented below.

5.1. Uncertainty and lack of awareness

Our data reveal that although researchers are familiar with ORD, they are uncertain of the data management practices of public data repositories (as ICSPR). This uncertainty is due to several reasons such as the lack of experience in ORD and the fact that many have not publicly shared and deposited their data in a data repository, as stated by R1:

‘There might be management systems...it might there is software available, I do not know enterprise system, I have no clue since I do not have the experience’. R1

While some journals and research funding institutions require making the research data open, this is not the norm and sharing research data is still mainly an optional choice and not a mandatory requirement for publication. Researchers thus were not motivated to familiarise themselves with available data repositories and their policies. One researcher stated:

‘it was not, to be honest, like made mandatory’. R6

This lack of interest in knowing more about data repositories and ORD practices was further associated with the fact that researchers run on a busy schedule,

‘researchers and professors, we have our schedule, it is generally busy’. R2

And as researchers are most of the time faculty members, they have other duties to attend to, and ORD is not part of these duties:

‘as a faculty member, responsibilities are more towards like doing some other duties. This seems to be something which is basically not the main part of the work’. R6

‘I didn’t really think of that too much because I was just focused on my research and on getting it done, and meeting deadlines, and doing a proper piece of work’. R7

Others acknowledged that the emerging nature of ORD is creating a sense of ambiguity and uncertainty. R2 argued that as ORD is an emerging phenomenon, different rules and practices are available creating confusion about what should and should not be done when sharing the data:

‘I don’t know, here, there is kind of a gray zone, that is not perfectly crystal clear. I mean, it’s an emerging movement,...we need to have some thinkers, I mean, researchers like you, who do research to come up with some best practices and rules, or some recommendations, okay, how we should proceed, how we should go’. R2

Accordingly, a clear and well-communicated legal framework for sharing research data is needed, as argued by R3:

‘So, the point is that there is no legal framework of how we are actually, as researchers, we’re going to share our data. So, all this is still not clear for us. It’s not clear at all. And it has not been communicated very clearly’. R3

5.2. Privacy concerns

Given that there was uncertainty about ORD practices, it was challenging for interviewed researchers to identify whether ORD entails any privacy concerns or possible privacy violations to research subjects. Nonetheless, they clarified their perspective from the general knowledge they have on ORD and open science paradigm.

Researchers emphasised the anonymity of the data explaining that as long as the data are anonymised, there should be no privacy issues. Equating privacy with anonymity was present in most of the interviews we conducted, as in the examples shown below:

‘I think if data is made perfectly 100% anonymous, later on, there is no worries in sharing it’. R2

‘all data, all the time, should not lead to identifying someone. So that’s important. And that’s what we do all the time. Make sure that we collect, it doesn’t lead to identify individuals’. R4

‘So I think that if you are using qualitative data, you need to make sure that the either the original organization that you are citing agrees to share publicly the data or you make all name anonymous’. R2

‘they don’t publish the whole data. So, this is really actually a portion of their data. And this has been adjusted and, you know, anonymized this and that. So, this is only a portion of their data, not the raw data. All the details. So, I think this is why they are not so concerned about who’s going to use, who’s going to look into it’. R9

Furthermore, some researchers believed that if the data are anonymised, then GDPR requirements are met:

‘Most of the data that I would collect, I would anonymize. It wouldn’t have personally identifying information. It couldn’t be linked to a specific person or individual So GDPR is addressed, I believe’. R10

Still, researchers displayed concerns over the possibility of re-identifying individuals and hence violating their privacy due to the use of artificial intelligence technology on research data and its ability to link the data to participants:

‘we have all these AI methods where you can just look at seeming anonymous data and generate some information that can perhaps be attributed to a person’. R12

AI technology, especially those based on large language models (e.g. ChatGPT) can be used to analyse ORD, which introduce privacy concerns:

‘I know...are already doing data analysis, qualitative and quantitative using ChatGPT or Bard...those models are learning about the respondents. If they have enough data the person can be known’. R6

Accordingly, researchers believed regulators must take these technologies into account when drafting or revising privacy laws:

‘I mean, they [regulators] are pretty sensitive about, protecting personal information...I think, they [regulators] still have a lot of ways to go with AI and use of AI’. R12

However, researchers agreed that if there are intentions to share the research data, the informed consent form must include a statement clarifying that the data collected might be shared in public repositories and used for purposes unknown to the researcher. Many researchers showed concerns and that they ‘would not feel comfortable’ to share the data if they did not take approval from research subjects on sharing their data, emphasising that the informed consent form acts as a moral and official agreement between the researcher and the research subjects. The following quotes explain this:

‘we tell them about how the data will be used, for how, for how long we’re going to keep it, and for what purpose? These all things should be clarified to the participant in the beginning...And if our plan is to share that data with third party or make it freely available, we need to disclose it. At that time’. R6

‘any good consent form fully informs research subjects about what will happen to their data’. R1

‘it is compulsory and it must be specific. You can’t ask for a general consent’. (R11)

‘I, personally, wouldn’t feel very comfortable doing so [sharing data without taking consent]’. R5

‘I think it shouldn’t be shared with the whole world because those people that we interviewed, so our participants, didn’t agree for their participation, so for their answers, for their contributions to be shared with just anyone. So I think...everyone should be informed of this’. R7

‘the agreement via this consent, both is moral and official agreement’. R3

Therefore, some see that sharing research data in public data repositories would not be conflicting with GDPR as long as participants were informed:

‘I think that the data protection principles are met, GDPR legislation does not prevent data collected by researchers from being archived and shared for other researchers to use. Researchers must ask consent from participants to reuse data in further research and to deposit it in an archive it’. R7

Despite agreeing on the importance of including the intentions of sharing the data in the consent form, researchers could not recall whether a clarifying statement about this is mentioned in the consent form they use. Researchers often rely on their universities’ guidelines and template for consent forms, which they believe comply with privacy regulations (e.g. GDPR) and policies that influence research data. In other words, researchers perceived that modifications to the consent form to comply with privacy regulations is the responsibility of their universities.

‘I use template from my university which was drafted by legal advisors. So my university or at least the legal service of my university knows what the latest laws require, what the GDPR requires, has specific templates for researchers in social science for instance’. R1

‘I’ve got to use different consent forms that comply with the university regulations, for instance, and the university regulations, it’s just GDPR. Yeah, it’s in conjunction with GDPR’. R3

‘but we as researchers, we need to operate under the rules, right, and the policies of the university’. R4

‘So everything here at least goes through the legal office, or the GDPR lawyer, who is also part of the legal office, of course’. R9

6. Discussion and implications

In this research, we sought to examine whether ORD paradigm may entail privacy violations to research participants. To achieve that, we first followed the heuristic steps suggested in privacy as contextual integrity theory [3]. We enriched our study by interviewing 12 researchers from European institutions to gain insights on their opinion of ORD and the possibility of privacy violations.

While prior studies examined factors that influence the wide-scale adoption of research data sharing and identified that concerns over confidentiality and data protection can inhibit academics from sharing their research data [29,30], scant research delved into these concerns. This research advances the current knowledge on ORD by offering in-depth understanding of the privacy implications of ORD and identifying the sources of tension between ORD policy and privacy protection regulation (GDPR).

Our research shows that according to privacy as contextual integrity theory, ORD may entail a violation to research participants’ informational privacy. At the heart of ORD paradigm is the provision of accessibility to research data to others to encourage knowledge creation [7,14,15]. This indicates that secondary use of data is promoted [6–8]. Nonetheless, using individuals’ information beyond the original purpose for which it was collecting, that is, secondary use, is often perceived as a violation of privacy that can have serious implications such as trust deterioration in the entity collecting the data [31]. Respecting privacy indicates respecting how information flows, between whom and for what purpose [4]. The GDPR adopts a control perspective on information privacy and hence it protects individuals’ right in deciding for themselves the particularities of how their information will be used. Although the GDPR offers exemptions to data collected for research purposes, these exemptions are conditioned on (1) having appropriate measures to protect the privacy of respondents’ data and (2) the expectations of respondents themselves. It is particularly the second condition that might drive privacy violation when considering research data sharing practices for data repositories, specifically ICSPR examined in this study. As mentioned earlier, public data repositories aim to encourage sharing research data. To achieve that, data repositories, and specifically ICSPR in our case, adopt a generic and simplistic approach to the language of the informed consent form, in an attempt to remove possible barriers to research subjects’ participation and consequently the possibilities of sharing research data. This generic language (e.g. ‘We may use the information you provide for future research studies’, ‘Your answers will be used *only* for the purposes of this study’) lacks transparency and accuracy, as

data in ORD might be used by other researchers than those who initially collected it, and used for other purposes than the one the research participants gave their consent on. Accordingly, data repositories guidelines for informed consent focus on protecting the confidentiality of research participants' data rather than their privacy.

It is worth mentioning that data repositories can have different guidelines and terms of use and hence the potential of research participants' data privacy violation might vary. For example, the public data repository Harvard Dataverse do not offer details about the privacy of research participants' data and how it should be protected, nonetheless, in its general terms of use, Harvard Dataverse clarifies that researchers upload the data under the restriction that any re-identification is not possible. Its terms of use specify:

'User Uploads must be void of all identifiable information, such that re-identification of any subjects from the amalgamation of the information available from all of the materials (across datasets and dataverses) uploaded under any one author and/or User should not be possible' [32].

Publishing identifiable information is allowed when all research subjects have given their explicit consent on the public sharing of their data in data repositories [32].

Therefore, different data repositories may have different approaches towards protecting research participants' data privacy; some offering detailed guidelines (e.g. ICSPR) while others leaving it more general (e.g. Harvard Dataverse). The researchers interviewed in this study expressed the need to adopt a more specific and transparent approach that is better oriented to protect privacy. They highlighted the importance to explicitly inform research participants of any intention to share data in public repositories as well as highlighting the implications of such sharing, for example, that their data might be used by other researchers for purposes that cannot be known in advance. This transparent approach better aligns with GDPR requirements in meeting participants' expectations, as well as the definition of privacy.

Understanding what privacy means and entails is crucial. While protecting individuals' privacy relates to protecting their rights on controlling the use of their data [4], our research reveals that researchers reduce privacy to anonymity. That is, as long as research participants' data are 'perfectly' anonymised, researchers believe there should be no privacy concerns. Following a reductionist approach to privacy where privacy equals anonymity not only places inappropriate constraints on the definition of privacy and hence how users' privacy should be protected, but it also introduces other challenges. First, researchers assume that perfect anonymity can be guaranteed despite the various evidence that re-identification is possible [33–35]. This is a critical issue to address as GDPR does not apply to anonymised data. Although guidelines on identifiability, anonymisation and pseudonymisation exist, it is apparent that researchers are not aware of these guidelines and what they need to do before sharing their research data. Second, research data might lose their value when contextual and identifiable information are anonymised or removed from the dataset [5,15] thereby defeating the goal of ORD.

Our qualitative study further revealed that researchers lack practical knowledge and awareness of the different aspects of ORD such as the type of data that can be shared, the processes of sharing and data repositories terms of use. A finding that is consistent with prior literature [36,37]. Indeed, experience and training on data sharing motivate researchers to share their research data [38] while ensuring the protection of research participants' data privacy. Moreover, our findings reveal that researchers are not familiar with the particularities of GDPR requirements, and many of them perceive that the responsibility to comply with GDPR lies within their universities.

7. Practical implications

Our research offers valuable practical implications. First, public data repositories might need to revise their guidelines and terms of use to better protect research participants' privacy and not just confidentiality. While it is understandable that data repositories might prefer the use of general language in the consent form to facilitate and encourage sharing research data, this should not come at the expense of protecting individuals' privacy. Privacy entails ensuring individuals' control over how their data are used. This indicates the need to be as transparent as possible when taking their consent. Such transparency is crucial to maintain their trust in research and research institutions. One way data repositories can balance their personal interest in data sharing with the need to comply with regulatory requirements on privacy is by changing their guidelines for informed consent to be better aligned with privacy requirements while also emphasising the confidentiality of participants' data.

Second, the use of AI tools in data collection and analysis places serious dangers on privacy as these tools enable re-identification. Clear guidelines on how to protect research data from re-identification possibilities should be provided by data repositories before sharing it. The European Union Artificial Intelligence Act can act as a basis for these repositories

to understand how AI might influence sharing research data and what legal obligations they might face. In addition, the GDPR should be revised to incorporate the threats AI tools might have on research data and how to mitigate them.

Third, researchers follow their university rules and requirements in terms of RDM practices. Consequently, researchers believe that academic institutions are the ones responsible for complying with privacy regulations. Accordingly, the legal department at universities must keep track with any changes or updates on privacy regulations and modify the university research practices to be aligned with these changes. One change can be updating the guidelines or template for informed consent to mention that data might be shared in public data repositories and used by other researchers for other research purposes when there is an intention to publicly sharing it. We aim to discuss our research findings with the legal department in our university and staff in the Deanship of Scientific Research to discuss the establishment of new guidelines related to protecting privacy in the context of sharing research data. In addition, it might be useful to organise a training and awareness session to researchers in our institution to demonstrate the importance of privacy-related issues in sharing research data and how to address them.

8. Limitations

Despite the valuable knowledge this research offers, it has certain limitations. The main limitation was the limited number of researchers who were interviewed. However, our aim of conducting the interviews was to enrich our findings, where the main purpose of this study is not to conduct interviews, rather to apply the heuristics steps of privacy as contextual integrity to answer whether sharing research data in public data repository entail privacy violation of research subjects. However, conducting another study with a larger number of researchers would yield a clearer picture of researchers' view in relation to ORD and privacy violation.

Another limitation to this study was the use of a single case repository, the ICPSR. We believe that different data repositories would have different methods of dealing with privacy violation, this may be further explored in future research. However, we still believe this study produces valuable findings related to sharing research data in public data repository and privacy violation.

9. Conclusion

We present an analysis of the tension between ORD and privacy regulations based on privacy as contextual integrity theory. According to the heuristics steps of privacy as contextual integrity and the GDPR requirements, we show that ORD might present a violation of research participants' privacy. Furthermore, we interviewed researchers from European institutions to explore their opinions on the privacy issues related to ORD and its alignment with the GDPR. Our interview data revealed that researchers conflate privacy and anonymity and lack knowledge of sharing research data practices. We present the tensions between public data repositories and privacy regulations and underscore the need to maintain a balance between the value of shared data while preserving the privacy of participants' data.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Laila Dahabiyeh  <https://orcid.org/0000-0002-5946-2256>

References

- [1] Borgman CL. The conundrum of sharing research data. *J Am Soc Inf Sci Technol* 2012; 63: 1059–1078.
- [2] Arza V and Fressoli M. Systematizing benefits of open science practices. *Inf Serv Use* 2017; 37: 463–474.
- [3] Nissenbaum H. Privacy as contextual integrity. *Wash L Rev* 2004; 79: 119–157.

- [4] Nissenbaum H. *Privacy in context: technology, policy, and the integrity of social life*. Stanford, CA: Stanford University Press, 2020.
- [5] Childs S, McLeod J, Lomas E, et al. Opening research data: issues and opportunities. *Rec Manag J* 2014; 24: 142–162.
- [6] Persic A, Beigel F, Hodson S, et al. *The time for open science is now*. UNESCO Science Report: The Race against Time for Smarter Development, 2021. London: UNESCO.
- [7] Longley Arthur P and Hearn L. Toward open research: a narrative review of the challenges and opportunities for open humanities. *J Commun* 2021; 71: 827–853.
- [8] Wessels B, Finn RL, Linde P, et al. Issues in the development of open access to research data. *Prometheus* 2014; 32: 49–66.
- [9] Open Data Institute. What makes data open? <https://theodi.org/insights/guides/what-makes-data-open/#:~:text=Open%20data%20is%20data%20that%20anyone%20can%20access%2C%20use%20and%20share.&text=The%20licence%20might%20also%20say,this%20is%20called%20share%2Dlike> (2013, accessed 20 September 2023).
- [10] European Commission. Facts and Figures for open research data, 2017, https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/open-science-monitor/facts-and-figures-open-research-data_en#funderspolicies
- [11] Martin-Melon R, Hernández-Pérez T and Martínez-Cardama S. Research data services (RDS) in Spanish academic libraries. *J Acad Libr* 2023; 49: 102732.
- [12] Mannocci A, Irrera O and Manghi P. Will open science change authorship for good? Towards a quantitative analysis. arXiv [preprint], 2022. DOI: 10.48550/arXiv.2207.03121.
- [13] Kwon S and Motohashi K. Incentive or disincentive for research data disclosure? A large-scale empirical analysis and implications for open science policy. *Int J Inf Manag* 2021; 60: 102371.
- [14] White T, Blok E and Calhoun VD. Data sharing and privacy issues in neuroimaging research: opportunities, obstacles, challenges, and monsters under the bed. *Hum Brain Mapp* 2022; 43: 278–291.
- [15] Hansson K and Dahlgren A. Open research data repositories: practices, norms, and metadata for sharing images. *J Assoc Inf Sci Technol* 2022; 73: 303–316.
- [16] Curty RG, Crowston K, Specht A, et al. Attitudes and norms affecting scientists' data reuse. *PLoS ONE* 2017; 12: e0189288.
- [17] Colavizza G, Hrynaskiewicz I, Staden I, et al. The citation advantage of linking publications to research data. *PLoS ONE* 2020; 15: e0230416.
- [18] Jao I, Kombe F, Mwalukore S, et al. Research stakeholders' views on benefits and challenges for public health research data sharing in Kenya: the importance of trust and social relations. *PLoS ONE* 2015; 10: e0135545.
- [19] Ni C, Cang LS, Gope P, et al. Data anonymization evaluation for big data and IoT environment. *Inf Sci* 2022; 605: 381–392.
- [20] Shvartzshnaider Y, Apthorpe N, Feamster N, et al. Going against the (appropriate) flow: a contextual integrity approach to privacy policy analysis. *Proceedings of the AAAI conference on human computation and crowdsourcing*, 2019, [https://nissenbaum.tech.cornell.edu/papers/Going%20Against%20the%20\(Appropriate\)%20Flow.pdf](https://nissenbaum.tech.cornell.edu/papers/Going%20Against%20the%20(Appropriate)%20Flow.pdf)
- [21] Zimmer M. Addressing conceptual gaps in big data research ethics: an application of contextual integrity. *Soc Media Soc* 2018; 4: 1–11.
- [22] Apthorpe N, Shvartzshnaider Y, Mathur A, et al. Discovering smart home internet of things privacy norms using contextual integrity. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2018; 2: 1–23.
- [23] Vitak J and Zimmer M. More than just privacy: using contextual integrity to evaluate the long-term risks from COVID-19 surveillance technologies. *Soc Media Soc* 2020; 6: 1–4.
- [24] Malkin N. Contextual integrity, explained: a more usable privacy definition. *IEEE Secur Priv* 2022; 21: 58–65.
- [25] ICPSR. *Guide to social science data preparation and archiving: best practice throughout the data lifecycle*. 6th ed. Ann Arbor, MI: ICPSR, 2023.
- [26] ICPSR. Recommended informed consent language for data sharing, <https://www.icpsr.umich.edu/web/pages/datamanagement/confidentiality/conf-language.html> (2023, accessed 10 November 2023).
- [27] Braun V and Clarke V. *Thematic analysis: a practical guide*. Thousand Oaks, CA: Sage, 2021.
- [28] Strauss A and Corbin J. *Basics of qualitative research techniques*. 2nd ed. Thousand Oaks, CA: Sage, 1998.
- [29] Zhu Y. Open-access policy and data-sharing practice in UK academia. *J Inf Sci* 2020; 46: 41–52.
- [30] Shmagun H, Shim J, Kim J, et al. Identifying key factors and actions: initial steps in the Open Science Policy Design and Implementation Process. *Journal of Information Science*. Epub ahead of print 31 October 2023. DOI: 10.1177/01655515231205496.
- [31] Martin K. The penalty for privacy violations: how privacy violations impact trust online. *J Bus Res* 2018; 82: 103–116.
- [32] Harvard Dataverse. General terms of use, <https://support.dataverse.harvard.edu/harvard-dataverse-general-terms-use> (2024, accessed 24 June 2024).
- [33] Douriez M, Doraiswamy H, Freire J, et al. Anonymizing nyc taxi data: does it matter. In: *2016 IEEE international conference on data science and advanced analytics (DSAA)*, Montreal, QC, Canada, 17–19 October 2016, pp. 140–148. New York: IEEE.
- [34] De Montjoye Y-A, Radaelli L, Singh VK, et al. Unique in the shopping mall: on the reidentifiability of credit card metadata. *Science* 2015; 347: 536–539.

- [35] Lomas N. Researchers spotlight the lie of ‘anonymous’ data, <https://techcrunch.com/2019/07/24/researchers-spotlight-the-lie-of-anonymous-data/> (2019, accessed 1 November 2023).
- [36] Chowdhury G, Boustany J, Kurbanoglu S, et al. Preparedness for research data sharing: a study of university researchers in three European countries. In: *Digital libraries: data, information, and knowledge for digital lives: 19th international conference on Asia-pacific digital libraries, ICADL 2017*, Bangkok, Thailand, 13–15 November 2017, pp. 104–116. Cham: Springer.
- [37] Steinhardt I, Bauer M, Wünsche H, et al. The connection of open science practices and the methodological approach of researchers. *Qual Quant* 2023; 57: 3621–3636.
- [38] Zuiderwijk A and Spiers H. Sharing and re-using open data: a case study of motivations in astrophysics. *Int J Inf Manag* 2019; 49: 228–241.