DESIDOC Journal of Library & Information Technology, Vol. 39, No. 6, Nov 2019, pp. 329-337, DOI : 10.14429/djlit.39.6.14807

Reusing Data: Technical and Ethical Challenges

Juan-José Boté* and Miquel Térmens

Departament de Biblioteconomia, Documentació i Comunicació Audiovisual
& Centre de Recerca en Informació, Comunicació i Cultura. Universitat de Barcelona.
C/ Melcior de Palau 140, 08014 ES, Barcelona, Spain
**E-mail: juanjo.botev@ub.edu*

ABSTRACT

Research centres, universities and public organisations create datasets that can be reused in research. Reusing data makes it possible to reproduce studies, generate new research questions and new knowledge, but it also gives rise to technical and ethical challenges. Part of these issues are repositories interoperability to accomplish FAIR principles or issues related to data privacy or anonymity. At the same time, funding institutions require that data management plans be submitted for grants, and research tends to be increasingly interdisciplinary. Interdisciplinarity may entail barriers for researchers to reuse data, such as a lack of skills to manipulate data, given that each discipline generates different types of data in different technical formats, often non-standardized. Additionally, the use of standards to validate data reuse and better metadata to find appropriate datasets seem necessary. This paper offers a review of the literature that addresses data reuse in terms of technical, ethical-related issues.

Keywords: Data research management; Data reuse; Ethical challenges; Research data management training; Library support services; FAIR principles, Literature review.

1. INTRODUCTION

Researchers create data using different methods during the research lifecycle. These data can be generated by any type of electronical object, such as sensors, mobile phones, smart meters, laboratory tools, voice interviews and electronic surveys, or manually, like interview transcriptions. These data are used for primary research and published in papers or research reports. Making them available for later reuse is important, but they may present technical and ethical challenges when being shared or reused. For instance, data may not have been designed for the purposed research or may not be structured in a way that is easy compatible[1].

Research data can be found in many places, such as public repositories, university repositories, webpages and even special collections[4]. Additionally, research data can include other types of data, such as metadata, texts, server logs and device specifications, that have not traditionally been considered research data[5].

The reuse of research data is defined as finding, processing and analysing someone else's datasets to create new knowledge. Pasquetto *et al*.[7] discuss the meaning of the term "reuse" when researchers reuse their own data. Data reuse comprises two completely different actions: data sharing and data publishing. Data sharing is when a researcher shares data with a peer, when the dataset is not necessarily publicly available. In this case, the data are likely to be trustworthy, and the information on how to reuse them, i.e. the documentation, allows them to be integrated into another research project more easily than in other situations. However, data sharing has different meanings in other research fields. In the field of health, for example, data sharing for the participants of any study involves informed consent to share their data under certain conditions, such as anonymity. Data publishing means that researchers deposit their research data in an institutional, general or specific repository to be reused at a later date. In this case, the data should be accompanied by detailed documentation on how to reuse them.

The different ways in which data can be reused has been discussed. Custers and Uršič[2] purposed a taxonomy based on reuse, repurposing and recontextualization, while Pasquetto[7] differentiates between independent reuse and combined reuse to build new models of data or explore new research questions.

Data reuse depends on factors such as users' IT (information and technology) skills and their ability to find, access and work with datasets[8]. It is unlikely that all data shared will be reused. Some fields still do not have a culture of data sharing. For example, few researchers publish new data in biodiversity databases[9]. In a study, conducted in Singapore, concerns regarding research data sharing included the misuse or misinterpretation of data[10].

One of the benefits of reusing datasets is the possibility of conducting secondary analysis. According to the legislation of some countries, research data can be used only for the project for which they were originally intended. For example, Morrow *et al*.[6] discuss the fact that data reuse can present some obstacles,

Received : 14 July 2019, Revised : 17 October 2019
Accepted : 13 November 2019, Online published : 11 December 2019

depending on the data-sharing conditions of the originator.

There are also social implications for the reuse of data. It was found that, with data collected by drones, it was possible to control protected biodiversity areas[71]. Through secondary analysis, it is also possible to analyse factors that influence productivity in agriculture[3], or benefits and risks of food composition, to improve public health initiatives[29]. In another study, it was found that analysing smartphone GPS (Global Positioning System) data could facilitate the planning and mitigation of future natural disasters[70].

In this paper, through literature review, we explore technical and ethical aspects of research data reuse. First, technical aspects are relevant to researchers since datasets can be misused. Every scientific discipline has a wide variety of digital formats, and research projects are increasingly interdisciplinary. In a scenario where different disciplines converge on a project, library support services are crucial to researchers, helping them to not only manipulate secondary datasets, but also produce new datasets. Technical aspects are therefore relevant to novice researchers who must acquire new skills relating to secondary analysis. Questions that are inherent to this matter include where and how to find research data and when to use them. Second, there are ethical challenges involved in secondary analysis. The level of data trustworthiness, as well matters related to informed consent, anonymity and the recognition of others' work, is relevant in any scientific fieldwork.

2. BENEFITS OF REUSING RESEARCH DATA

There are several benefits of reusing research data, including cost effectiveness and efficiency, an increased sense of community, greater transparency and clarity of research, recognition of data ownership and an increase in citations. It enables researchers to save time on data collection, but they must learn the methodology involved in obtaining data or obtaining all possible documentation on how data was collected[11]. Additionally, it allows both researchers and institutions that store data to enhance research data management.

In economic terms, reusing data allows researchers and institutions to save money, but institutional repositories need a return on investment when they share data; some institutions may be reluctant to invest in sharing data because they may not obtain an immediate return[13]. Based on this point, a mathematical model was created to calculate the break-even point for time spent sharing data in a scientific community versus time gained by reusing data[12].

In terms of research, it gives rise to new challenges in the research cycle, such as the study design phase and during and after the research. For instance, new hypotheses and new research questions can be developed and new interpretations of data can be made[11]. Moreover, Pasquetto *et al.* [7] state that data sharing makes it possible to reproduce research and advance science and innovation. They also explore distinctions between "use" and "reuse" and define the former as the processing of primary data by an individual or team for the purposes of a project, while reuse is when someone other than the originator uses the data. Therefore, the consistent citation of datasets would increase their dissemination. It is important to understand that experimenting with research involves thinking not only about how to manipulate the data, but about how to capture the right metadata and how to share the data in a way that allows others to reuse it later[16]. Metadata allow data to be found more easily, so systems could automatically add metadata to facilitate searches[5]. Finally, reusing data makes it possible to create a new dataset and publish it for future use, thereby avoiding unnecessary duplication of data collection work[13].

It also presents advantages with respect to skills, as it allows researchers to become better research data managers and improves data management in scientific disciplines internationally[13]. Said skills are essential for researchers to fully engage with the whole process of the research life cycle[17]. Nevertheless, a study that analysed American and Canadian repositories found that research data services were rarely implemented[18].

Regarding qualitative secondary analysis, there are two types of data that can be reused: others' and one's own ("auto-data"). And although the secondary analysis of auto-data seems to be neglected in the literature, it can offer a number of benefits, such as helping to increase the sample size in a determined context and protecting the privacy and confidentiality of participants[19].

3. TECHNICAL CHALLENGES OF REUSING DATA

Obtaining secondary research data is not technically straightforward and several questions need to be answered before a dataset can be reused. To be reused, research data should meet the FAIR (findable, accessible, interoperable and reusable) principles proposed[20] and subsequently adopted by the European Commission[21].

As explained above, research data can be found when data are shared by a peer or published in a repository. The first question is, where can the data be found if they have not been shared? Most experienced researchers are familiar with public and institutional repositories and would probably check there first. However, this is not the case with novice researchers, who have more barriers to overcome. The term "findable" implies that a researcher can easily find the dataset. When datasets or other information are published in an institutional repository, they are usually indexed in scholarly search engines such as Google Scholar and Microsoft Academics. However, in the case of datasets, Google also has a specific search engine, Google Dataset, which is currently in its beta version. Specific repositories in fields such as genetics and biomedicine are also indexed when internal permissions are granted to crawlers. Additionally, repositories have been found to present a number of challenges for users with respect to video datasets, such as advanced tools, data management and interoperable collaboration[8].

The second question concerns accessibility. Data are created in a wide range of digital formats and good IT skills are sometimes necessary. For instance, in generalists' repositories it is possible to find heterogenous data formats that require different tools and skills to reuse them[22]. Data in institutional and public repositories are either created digitally or digitised. In the case of digitised datasets, these are not always accessible

through commons tools, meaning that one challenge is how to manipulate the data with the adequate software, since not all universities or researchers know about or use the same tools and software. For example, the use of different statistical software packages can lead to misuse of data because of the format of the data. However, it is proposed that, to accomplish FAIR principles, data should be a in non-proprietary format such as CSV (comma separated value) or TSV (tab-separated values) to better facilitate reuse[23].

Dealing with issues such as the recognition of digital formats can also be problematic. As there are different digital format types, data may take different forms – for example, spoken data, images and videos – and may not always be standardised. With respect to spoken datasets, an open-source software was proposed to visualise large prosody spoken corpora data[24]. In another study, an open-source software to integrate phonetic analysis was proposed[25].

In several research fields, datasets must be anonymised, which raises the challenge of codification and dataset documentation. While anonymisation tools to apply deep learning techniques seem not to reduce the accuracy of data analysis[26], there is no standard for codifying data, and this presents a barrier for researchers who want to reuse data. For example, limitations and technical problems have been encountered with codified files in government data because of anonymisation, though clear data policies could be a solution to this problem[14]. In fields that use videos of participants, tools to support anonymity may not obscure all aspects of individuals and may leave identifiable characteristics, thereby compromising peoples' privacy and confidentiality[27]. However, obscuring people in videos may erase contextual information and study results can become less trustworthy.

There are several occasions when documentation is a priority – for example, if a researcher leaves a project. In this case, it becomes necessary to ensure that detailed documentation is available to check the dataset's accuracy, providing details to other researchers about the context, relevancy and trustworthiness of the data[28].

Accessibility also relies on being able to use old data, which can be accomplished through adequate digital preservation policies. Digital preservation is essential in repositories, not only to maintain digital data, but also to make data accessible through adequate processes, make research economically sustainable and ensure a return on investment. This can be achieved, for instance, by standardising and unifying file formats to provide better user support[30]. It is important for researchers to know in advance that their data will be maintained for a long period, not only for reuse, but for citation purposes as well. Consequently, migration and format control seem to be essential operations for data repositories. Most reputable repositories include preservation in their terms of use in addition to their research data management policies, but this is not always the case[18].

Interoperability cannot be possible in the absence of metadata. For instance, a solution proposed in the field of earth science was to provide a way to check the understandability and usability of data[31]. With respect to reviewing datasets, human-readable metadata are critical because peer review is not going to be carried out by machines any time soon. Therefore, there should be more integration between repositories[32]. This would not only facilitate interoperability but would also increase the use of datasets. In the field of health, it was found that standardised, structured, electronic health records were under development, and this should lead to reliable information and offer interoperability as a benefit in secondary analysis[33].

Metadata in datasets facilitate not only interoperability but also gives context to the information, such as how it was collected and how it should be treated. Thus, contextual information refers to the set of interrelated environmental conditions in which data are produced[22]. Depending on the research discipline, the challenges associated with reuse result from the wide variety of variables, which are collected in different ways[34].

Data Documentation Initiative (DDI, https://www.ddialliance.org/) is a standard for describing data generated by surveys and other observation methods; poor data documentation practices can lead to misuse or misinterpretation of the data.

By contrast, good documentation practices and the use of standards facilitates interoperability among different types of datasets. In addition, it would give rise to trustworthiness, transparency and verifiability. For instance, in a study in the field of phenomics, a repository was built that links phenomic, genomic and genetics for plants and their pathogens that, with the use international standards, allows for not only the reuse of datasets but also interoperability with other repositories[36]. It is important to be able to identify the provenance of the data through good documentation habits, and consistently documented provenance and context in all disciplines requires a joint effort.

During an analysis of food consumption apps that allow data to be reused, a major lack of documentation was found regarding data export, terms of use and privacy policies. Consequently, the apps did not comply with two of the FAIR principles: accessibility and interoperability. Furthermore, food consumption information could be considered personal data. Therefore, this would present a challenge in Europe, given that General Data Protection Regulation (GDPR) would apply[38]. Another study related to commercial pig farming found that analysing secondary data was a challenge since there were limitations related to the combined use of several data sources. Missing values made it impossible to conduct the study, and discrepancies in public databases made it necessary to collect information from farmers, which was time consuming[39].

Nevertheless, while the FAIR principles may pose restrictions in some fields, the nature of data means that some security is required. For instance, in human genomics, data cannot simply be made available without some form of access control[35]. In education, the privacy of subjects in videos requires access control in order to protect participants from personal or economic harm, or harm related to exposure of their identity[23].

3.1 Best Practices to Solve Technical Challenges

To solve technical problems, several authors have recommended best practices not only to share data more

effectively, but also to address other matters affecting data, such as research reproducibility. For example, in the field of health, it was found that video datasets were not published or linked to scientific papers, and most of the studies analysed could not be reproduced[40]. In such a scenario, which presents a lack of published datasets, the conclusions cannot be considered valid[31].

It was proposed, for instance, that research data be accompanied by a data publication process (as with scientific papers) and be subject to editorial quality control and an independent peer-review system[9]. This would also create an incentive to increase visibility, and authorship recognition would increase citation rates. However, sharing primary data requires effort and is sometimes considered a waste of time by researchers[41]. Given that sharing data does not count towards an academic career, there seems to be little incentive for researchers to publish their research data.

Making data and software reusable and documenting the provenance of computational results were proposed as best practices in geoscience, in addition to those recommended by organisations such as the Research Data Alliance[42].

Tracking the impact of the research data would reveal whether or not they were completely reused, and university libraries could use systems to identify patterns that reveal which data are reused most often[43]. These systems would track where they are published and identify which collections of data were reused most. Another method that has been proposed is to measure data from downloads of data held at repositories[44].

With respect to qualitative secondary analysis, there are several questions related to best practices. There is a need for informed consent that states that data will be reused; however, re-contacting participants for secondary analysis would also represent an advantage[19].

In ethnographic fieldwork and in interviews, qualitative data obtained and later transcribed may not convey the same meaning as when they were obtained. Additionally, reconstructing the situation when the transcriptions are computer-processed can pose problems (in the case of whispering, for instance) but using memos for clarification would partially help[45].

Regarding quality, validity and reliability, it is advisable to ensure that data are accurate, with no typographical errors or incomplete sentences, and that focus-group sessions are accurately transcribed and transcription documents time-stamped. Additionally, accessing information on how study participants were selected and recruited would be helpful[46].

4. ETHICAL CHALLENGES OF REUSING DATA

There are several ethical matters concerning data reuse, as each scientific discipline collects its own data and ethical concerns tend to emerge by the time the research data is reused. The first concern relates to the level of trustworthiness of the data with respect to provenance and reliability. The second relates to anonymity, privacy and confidentiality. The third relates to recognition of authorship when reusing others' data. Finally, data licencing is essential when reusing research data.

4.1 Level of Data Trustworthiness

The level of trustworthiness of data is an important factor for researchers who want to reuse research data. Scientists must be able to trust the reliability of the data they are going to use, but no standards for validating the reuse of qualitative or quantitative data have yet been established, and it seems that researchers are required to put their trust in many aspects of the data they reuse. However, the level of trustworthiness of the data is not always clear, especially if a dataset comes from an unknown repository. In the field of cancer epidemiology, for example, small datasets are currently used and reused in new analyses but are difficult to find because they are rarely deposited in repositories and are instead published on trusted social networks[75]. Evidence for the trustworthiness of the data is therefore often limited to evaluations of the data and ethics from the original study[48].

Once a dataset has been downloaded, it usually needs to be cleaned, integrated and analysed with other data. However, this integration requires a level of trust in the data so as not to obtain a biased result in the new research. For instance, a study on the reuse of quantitative data concluded that researchers trusting data was not a simple process, and that there are different levels of trustworthiness[48]. However, a lack of trustworthiness was not a factor in the failure to reuse data[49].

Trustworthiness refers also to whether the process used to create the data is credible, whether the data is consistently generated, and whether rigorous analysis methodologies are followed to ensure that the data is credible for reuse[50]. Reusing auto-data also offers the benefit of trustworthiness, because the context, the data collection method and the publication details are known[19]. In Kenya, social relations were found to be relevant to trusting shared data[51].

In clinical trials, sharing data for re-examination and replication of analysis ensures that important results, that are either intentionally hidden or inadvertently omitted, are revealed. Not sharing data prevents society from benefitting from clinical trials and may cause harm to participants because of undiscovered insights[52]. Therefore, in trials, the reuse of data is essential to the credibility, reliability and trustworthiness of the research.

4.2 Informed Consent

In secondary analysis, one of the main issues expressed throughout the scientific literature was informed consent. Informed consent usually relates to data sensitivity, privacy, confidentiality, anonymity, sharing and subsequent reuse. It is a communication process in which study participants accept or refuse to allow their data to be used in research, usually anonymously[54]. In the field of health, for example, it has been suggested that data entered in electronic health records should not be used for research without patients' consent[55]. In the case of clinical trials, it was proposed that broad consent to secondary use of data should become standard procedure, especially in Europe, with the introduction of the GDPR[53] . It was also proposed that true informed consent should indicate a specific research purpose, where the use of data is limited to one study. However, this would decrease the utility of data[56].

Surmiak[57] reported that, to obtain informed consent, researchers communicate where the data is stored and how it can be accessed, but this entails a rigorous process to ensure data anonymity so as not to expose participants to harm or discomfort. Zook *et al.*[58] present a set of rules for addressing issues such as potential harm, privacy and de-identification of data. For instance, they propose that a code of conduct be established for the research community. The best way to achieve informed consent is to inform participants about the purpose of the research and to explain how their data may be used in the future[15].

Any user can withdraw informed consent and even modify what kinds of data can and cannot be shared. Such situations may present challenges, not only with respect to reusing data, but also when it comes to de-identifying data and making the corresponding changes. Use of the sample is thus limited while data are removed from the study. Consequently, broad consent is widely used for biobanks because their aim is to use data samples in several research projects and donors are deprived of the possibility of withdrawing their consent[56]. A study conducted in Thailand regarding broad consent suggested that participants should be informed that consent to data sharing will not result in any re-contact[59].

In the case of qualitative secondary research, it seems that there is a lack of practical guidance, and it is vital in this type of research to explicitly outline how informed consent is sought to avoid harm[60]. In Europe, seven different types of informed consent were found in the literature: explicit consent, dynamic consent, individual consent, meta consent, consent for contact, consent agreement with GP (general practitioner) and opt out[61]. Moreover, re-contacting participants in a qualitative study could cause psychological, social or other harm[19].

4.3 Anonymity

Data anonymity is essential in secondary analysis. In this respect, El Eman *et al*.[60] reported that it is expected that anonymised data is only used for purposes that are legitimate, which is explicit in the context of the European Union. In addition, Sanchez *et al*.[64] reported that it is possible to anonymise data using techniques well stablished in the literature. However, in the field of asylum claims, for example, the risks of dual-use may arise even when data are anonymised[62]. Curty *et al.*[63] states that, in the field of social sciences, where human interaction is involved in data collection, there are many ethical concerns about sharing and reusing data. In the case of sharing, qualitative data, including consent, requires precise handling, and the authors suggest that social scientists be provided with guidance on data reuse practices.

In the field of ecology, there are no ethical norms but, as a general rule, locations that may endanger sensitive species are not published and data that could harm people are not shared[16]. However, in online research, anonymity may seem difficult to ensure, and not all users may want to preserve their anonymity. In addition, quoting online text may require permission as it may be tracked. Consequently, to reduce discoverability, some data can be summarized without losing meaning and other details altered or removed[65].

4.4 Recognising Authorship

Ethical practices entail recognising authorship, which involves not only recognising others' work, but also the possibility of collaboration with others. One of the issues that arises when reusing data is citation – in other words, recognising the authorship of others' work. Therefore, it was suggested that scientists would widely share their data as long as they were paid in form of reputation[37].

Depending on the importance of the data and the new knowledge generated, the creator of the data is likely to be offered co-authorship of a paper. One of the reasons for this is that the data's creator knows the context in which the data have been created, especially in qualitative studies. Bierer *et al*. proposed a system of recognition where data generators could be identified in a standardised way, differentiating from the authors of a peer-reviewed journal article, because data authors are responsible for the integrity of the data[74]

Transparency would increase if there were an automatic qualitative system to peer review research. Additionally, it also would avoid the need for researchers to clean the data[9].

Other aspects that may influence the recognition of authorship are the use of illicit datasets and copyright licences. One study, in the field of computer science, suggested that obtaining illicit datasets could be advantageous and would require fewer resources than collecting data from scratch, but it may pose questions about the use of these data[66]. In addition, it may lead to legal issues, depending on the country where the data is stored.

Related to copyright, another relevant aspect is understanding intellectual property rights related to copyright licences and the use of Creative Commons licences to ensure that others properly attribute credit for any dataset that is reused[3]. However, in some countries, the reuse of datasets maybe limited by laws. Slavnic[67] reported that, in Sweden, there is a paradox, as there is infrastructure in place for arching and reusing data, but it is illegal to reuse them for projects other than those for which they were originally obtained. In countries like Germany, there are also challenges relating to copyright and legality[68]. Therefore, datasets might need to be subject to restricted access requiring patents, or other specific situations.

5. DISCUSSION

The findings of this study explain that there are benefits to reusing research data, but there also several challenges and barriers to face. Digital preservation is essential to data reuse and facilitates access to the data in the long term. Therefore, repositories in higher education institutions and research centres are essential for digital preservation, as well as data sharing and reuse. However, there are still technical and ethical questions to face, and, in our opinion, barriers to overcome to fully be able to profit from data sharing and reuse. In addition, sharing and reusing data for secondary analysis is not common in all research fields[69].

Data reuse facilitates secondary analysis and research reproduction, but the reuse of data is dependent upon a diverse set of factors, such as good data documentation, interoperability

among repositories, data anonymity and, especially, the trustworthiness of the data. While often considered to be a waste of time[41], documenting data ensures that, before the data is shared, additional information in the form of reports or metadata is detailed to aid in the reuse of the data by other researchers. Nevertheless, we think that auditing data and documentation is also necessary to guarantee not only data validity and reliability but also the reputations of researchers. Auditing data could be done through different methods, such as specialised software or a peer-review process[9]. Interoperability seems a common attribute of specialized repositories, but in generalist repositories the reuse of data seems difficult due to the quantity and diversity of information. This makes it difficult not only reuse data but also to assess its level of trustworthiness due to the lack of standards[48]. Consequently, researchers sharing data via specialized repositories seems to be the best option. Data anonymity is relevant because it facilitates data reuse, but, in some fields, it may lead to the context in which the data were collected being lost, and participants may not want to remain anonymous[65]. Regarding privacy and data protection, it seems there is a barrier to reusing data due to differing legislation between countries, as could happen in the case of data held by the European Union[73]; it may cause concern if researchers from one country reuse data from another country, if the data need to be reused for a purpose other than what was originally intended, or if they need to be recontextualised.

The benefits of reusing data seem clear to researchers; saving money and time are possible benefits, as well as increasing the credibility of the original researcher and providing consistency among research. In addition, reusing data can lead to the creation of new datasets and new knowledge, but this is only possible when later reuse processes are well documented and follow good digital preservation practises.

Another benefit for researchers is authorship recognition and an increase in citations. Authorship recognition allows researchers to be recognized for their work, but there are still ethical issues to solve, such as data licensing, data provenance and collecting[66].

An improved system of tracking data was established[43], but this is not standardised or common in research. In addition, there was a proposal to recognize data creators in scientific papers[74]. We believe that this is an area where it is necessary to find solutions, and that further research is needed.

Society has also benefitted from the reuse of research data; natural disaster prevention[70], climate change mitigation[71] and mobility enhancement in urban environments[72] have all benefitted from data reuse. In clinical trials[65], where data sharing and reuse is common, it can improve the reliability of research leading to new discoveries in health which benefit the population. In addition, we think that there are other disciplines that can benefit society from the reuse of data, such as archaeology and economics.

6. CONCLUSIONS

In this study, we focused reviewing technical and ethical issues that data sharing and data reuse may present. Relating to technical issues, interoperability of repositories, data validity and reliability are likely the main concerns. The absence of standards in some areas, such as data codification, and the lack of standardised technical formats for data are also issues limiting data reuse. This means that generalist repositories often do not adhere to FAIR principles with respect to reusing both quantitative and qualitative research data. To start solving this issue, researchers should receive training in the data creation process, with an emphasis on the fact that data documentation should be carried out in the early stages[68] of the research, not only to facilitate subsequent reuse, but to ensure that they are digitally preserved long-term. Relating to ethical issues, informed consent seems necessary in some disciplines but not in others, and a better consensus should be found, especially when the reuse of data is among researchers from different countries, where legislation is sometimes completely different. Aspects other than informed consent, such as data anonymity, are also important, with an emphasis on qualitative research. Our further study will be focused on data research, auditing methodologies to provide reliability and validity to research data.

REFERENCES

1. Milne, D. & Watling, D. Big data and understanding change in the context of planning transport systems. *J. Transp. Geogr*., 2019, **76**, 235–244. doi: 10.1016/j.jtrangeo.2017.11.004.
2. Custers, B. & Uršič, H. Big data and data reuse: A taxonomy of data reuse for balancing big data benefits and personal data protection. *Int. Data Privacy Law*. 2016, **6**(1), 4–15. doi: 10.1093/idpl/ipv028.
3. Charamba, V.; Thomas, B. & Charamba, B. Relative importance analysis of the factors influencing maize productivity at Olushandja and Etunda irrigation schemes of Namibia: A secondary analysis of data from farm household survey. 2017. http://repository.unam.edu.na/handle/11070/2193 (accessed on 10 June 2019).
4. Boté, J. Dataset management as a special collection. *Collect. Manage.* 2019, **4**(2-4), 259-276. doi: 10.1080/01462679.2019.1586613.
5. Gregory, K.; Cousijn, H.; Groth, P.; Scharnhorst, A. & Wyatt, S. Understanding data search as a socio-technical practice. *J. Inf. Sci*., 2019. doi: 10.1177/0165551519837182.
6. Morrow, V.; Boddy, J. & Lamb, R. The ethics of secondary data analysis: Learning from the experience of sharing qualitative data from young people and their families in an international study of childhood poverty. 2014. http://sro.sussex.ac.uk/id/eprint/49123/ (accessed on 5 June 2019).
7. Pasquetto, I.; Randles, B. & Borgman, C. On the reuse of scientific data. *Data Sci. J.,* 2017, **16**(8). doi: 10.5334/dsj-2017-008.
8. Frank, R.D.; Suzuka, K. & Yakel, E. Examining the reuse of qualitative research data: Digital video in education. *In* Archiving Conference, Archiving 2016 Final Program and Proceedings, pp. 146-151(6). doi: 10.2352/issn.2168-3204.2016.1.0.146.
9. Costello, M.J.; Michener, W.K.; Gahegan, M.; Zhang, Z. & Bourne, P.E. Biodiversity data should be published,

cited, and peer reviewed. *Trends Ecol. Evol.* 2013, **28**(8), 454–461.
doi: 10.1016/j.tree.2013.05.002.

10. Majid, S.; Foo, S. & Zhang, X. Research data management by academics and researchers: perceptions, knowledge and practices. *In* maturity and innovation in digital libraries, edited by Dobreva, M.; Hinze, A. & Žumer, M. Springer International Publishing. 2018. 166–178
11. Johnston, M.P. Secondary data analysis: A method of which the time has come. *Qual. Quant. Methods in Libr.* 2017, **3**(3), 619–626. http://www.qqml-journal.net/index.php/qqml/article/view/169. (accessed on 14 June 2019).
12. Pronk, T.E. The time efficiency gain in sharing and reuse of research data. *Data Sci. J.,* 2019*,* **18**(1), 1-8.
doi: 10.5334/dsj-2019-010
13. Figueiredo, A.S. Data sharing: Convert Challenges into Opportunities. *Front. Public Health.,* 2017, **5**.
doi: 10.3389/fpubh.2017.00327.
14. Attard, J.; Orlandi, F.; Scerri, S. & Auer, S. A systematic review of open government data initiatives. *Gov. Inf. Q.* 2015, **32**(4), 399–418.
doi: 10.1016/j.giq.2015.07.006.
15. Van den Eynden. Informed consent for data sharing and reuse. Creating shareable research data: Managing and archiving social science research data*.* 2017*.* https://ukdataservice.ac.uk/media/605026/2017-11-28_consent_final_-pdf.pdf
16. Hampton, S.; Anderson, S.; Bagby, S. *et al.* The tao of open science for ecology. *Ecosphere.* 2015, **6**(7).
doi: 10.1890/ES14-00402.1.
17. Shorish, Y. Data information literacy and undergraduates: A critical competency. *College Undergrad. Libr*., 2015, **22**(1), 97–106.
doi: 10.1080/10691316.2015.1001246.
18. Tripathi, M.; Shukla, A. & Sonkar, S.K. Research data management practices in university libraries: A study. *DESIDOC J. Lib. Inf. Tech.,* **37**(6), 2017, 417–424.
doi: 10.14429/djlit.37.6.11336.
19. Watters, E.C.; Cumming, S. & Caragata, L. The lone mother resilience project: A qualitative secondary analysis. *Forum Qual. Sozialforschung / Forum: Qual. Soc. Res.*, 2018, **19**(2).
doi: 10.17169/fqs-19.2.2863.
20. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J. *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci. Data,* 2016*,* **3**. https://www.nature.com/articles/sdata201618 (accessed on 10 April 2019).
21. Data management - H2020 online manual. http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm (accessed on 10 June 2019).
22. Assante, M.; Candela, L.; Castelli, D. & Tani, A. Are scientific data repositories coping with research data publishing? *Data Sci. J*., 2016, **15**(6).
doi: 10.5334/dsj-2016-006.
23. Jacob, D. FAIR principles, a new opportunity to improve the data lifecycle. in Proceedings of ado2019: Journée thématique sur les autorités de données (Pascal Dayre - CNRS ENSEEIHT-IRIT, 2019). https://hal.archives-ouvertes.fr/hal-02070883/ (accessed on 10 June 2019).
24. Öktem, A.; Farrús, M. & Wanner, L. Prosograph: A tool for prosody visualisation of large speech corpora. *In* Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017). 2017. Stockholm, Sweden. http://hdl.handle.net/10230/32719 (accessed on 10 June 2019).
25. McAuliffe, M. et al. ISCAN: A system for integrated phonetic analyses across speech corpora. 2019. http://eprints.gla.ac.uk/183719/ (accessed on 10 June 2019).
26. Niimi, A. Study on data anonymization for deep learning. in recent trends and future technology in applied intelligence. *In* Proceedings of International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. 2018. pp. 762-767.
27. Frank, R.D.; Tyler, A.R.B.; Gault, A.; Suzuka, K. & Yakel, E. Privacy concerns in qualitative video data reuse. *Int. J. Digital Curation,* 2019, **13**(1), 47–72 (2019).
doi: 10.2218/ijdc.v13i1.492.
28. Faniel, I.M.; Kriesberg, A. & Yakel, E. Social scientists' satisfaction with data reuse. *J. Assn. Inf. Sci. Tec.,* 2016, **67**(6), 1404–1416.
doi: 10.1002/asi.23480.
29. Neufingerl, N. et al. Intake of essential fatty acids in Indonesian children: Secondary analysis of data from a nationally representative survey. *Br. J. Nutr.* 2016, **115**(4), 687–693.
doi: 10.1017/S0007114515004845.
30. Termens, M.; Ribera, M. & Locher, A. An analysis of file format control in institutional repositories. *Libr. Hi. Tech.,* 2015*,* **33**(2), 162–174.
doi: 10.1108/LHT-10-2014-0098.
31. Callaghan, S. Data without peer: Examples of data peer review in the earth sciences. *D-Lib Magazine,* 2015, **21**(1-2).
doi: 10.1045/january2015-callaghan.
32. Tenopir C.; Dalton, E.D.; Allard S.; Frame M.; Pjesivac I.; Birch B. *et al.* Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS ONE,* 2015*,* **10**(8):e0134826.
doi: 10.1371/journal.pone.0134826.
33. Vuokko, R.; Mäkelä-Bengs, P.; Hyppönen, H.; Lindqvist, M. & Doupi, P. Impacts of structuring the electronic health record: Results of a systematic literature review from the perspective of secondary use of patient data. *Int. J. Med. Inf.,* 2017*,* **97**, 293–303.
doi: 10.1016/j.ijmedinf.2016.10.004.
34. Yoon, A.; Jeng, W.; Curty, R. & Murillo, A. In between data sharing and reuse: Shareability, availability and reusability in diverse contexts: In between data sharing and reuse: Shareability, availability and reusability in diverse contexts. *Proc. Assoc. Info. Sci. Tech.,* 2017, **54**(1), 606–609.
doi: 10.1002/pra2.2017.14505401085.
35. Boeckhout, M.; Zielhuis, G.A. & Bredenoord, A.L.

The FAIR guiding principles for data stewardship: fair enough?. *Eur. J. Hum. Genet.,* 2018*,* **26**(7), 931-936. doi: 10.1038/s41431-018-0160-0.

36. Pommier, C. et al. Applying FAIR principles to plant phenotypic data management in GnpIS. Plant phenomics. 2019. doi: 10.34133/2019/1671403.
37. Fecher, B.; Friesike, S.; Hebing, M.; Linek, S. & Sauermann, A. A reputation economy: Results from an empirical survey on academic data sharing. 2015. https://arxiv.org/ftp/arxiv/papers/1503/1503.00481.pdf(accessed on 10 June 2019).
38. Maringer, M.; Van't Veer, P.; Klepacz, N. *et al.* User-documented food consumption data from publicly available apps: An analysis of opportunities and challenges for nutrition research. *Nutr. J.,* 2018, **17**(59). doi: 10.1186/s12937-018-0366-6.
39. Pandolfi, F.; Edwards, S. A.; Maes, D. & Kyriazakis, I. Connecting different data sources to assess the interconnections between biosecurity, health, welfare, and performance in commercial pig farms in Great Britain. *Front. Vet. Sci.,* 2018, **5**. doi: 10.3389/fvets.2018.00041.
40. Boté, J. Lack of standards in evaluating YouTube health videos. *Revista Cubana de Información en Ciencias de la Salud,* 2019, **30**(2). http://www.acimed.sld.cu/index.php/acimed/article/view/1357 (accessed on 5 June 2019).
41. Fecher, B.; Friesike, S. & Hebing, M. What drives academic data sharing?. *PLoS ONE*, 2015, **10**(2): e0118053. doi: 10.1371/journal.pone.0118053.
42. Gil, Y.; David, C.H.; Demir, I. *et al.* Toward the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance: geoscience paper of the future. *Earth Space Sci.,* 2016*,* **3**(10), 388–415. doi: 10.1002/2015EA000136.
43. Ball, A. & Duke, M. How to track the impact of research data with metrics (DCC How-to Guides). 2015. http://www.dcc.ac.uk/resources/how-guides/track-data-impact-metrics (accessed on 5 June 2019).
44. Bishop, L. & Kuula-Luumi, A. Revisiting qualitative data reuse: a decade on. *SAGE Open*, 2017, 1-15. doi: 10.1177/2158244016685136.
45. Stuckey, H.L. The second step in data analysis: Coding qualitative research data. *J. Soc. Health Diabetes.* 2015, **3**(1), 7–10. doi: 10.4103/2321-0656.140875.
46. Sherif, V. Evaluating preexisting qualitative research data for secondary analysis. *Forum Qual. Sozialforschung / Forum: Qua. Soc. Res.,* 2018, **19**(2). doi: 10.17169/fqs-19.2.2821.
47. El Emam, K.; Rodgers, S. & Malin, B. Anonymising and sharing individual patient data. *B.M.J.*, 2015, **350**, h1139. doi: 10.1136/bmj.h1139.
48. Yoon, A. Data reusers' trust development. *J. Assoc. Inf. Sci. Technol.,* 2017, **68**(4), 946–956. doi: 10.1002/asi.23730.
49. Curty, R.G.; Crowston, K.; Specht, A.; Grant, B.W. & Dalton, E.D. Attitudes and norms affecting scientists' data reuse. *PLoS ONE,* 2017, **12**(2): e0189288. doi: 10.1371/journal.pone.0189288.
50. Nowell, L.S.; Norris, J.M.; White, D.E. & Moules, N.J. Thematic analysis: Striving to meet the trustworthiness criteria. *Int. J. Qual. Methods*. 2017, **16**. doi: 10.1177/1609406917733847.
51. 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**(9): e0135545. doi: 10.1371/journal.pone.0135545.
52. Bauchner, H.; Golub, R.M. & Fontanarosa, P.B. Data sharing: An ethical and scientific imperative. *JAMA*. 2016, **315**(12):1238-1240. doi: 10.1001/jama.2016.2420.
53. Ohmann, C. et al. Sharing and reuse of individual participant data from clinical trials: Principles and recommendations. *BMJ Open*. 2017,7:e018647. doi: 10.1136/bmjopen-2017-018647.
54. Grady, C. Enduring and emerging challenges of informed consent. *N. Engl. J. Med.,* 2015*,* **372**, 855–862. doi: 10.1056/NEJMra1411250.
55. Yen, J.C.; Chiu, W.T.; Chu, S.F. & Hsu, M.H. Secondary use of health data. *J. Formosan Med. Assoc.,* 2016, **115**(3), 137–138. doi: 10.1016/j.jfma.2015.03.006.
56. Sariyar, M.; Schluender, I.; Smee, C. & Suhr, S. Sharing and reuse of sensitive data and samples: Supporting researchers in identifying ethical and legal requirements. *Biopreserv. Biobanking,* 2015, **13**(4). doi: 10.1089/bio.2015.0014.
57. Surmiak, A.D. Confidentiality in qualitative research involving vulnerable participants: Researchers' perspectives. *Forum Qual. Sozialforschung / Forum: Qual. Soc. Res*., 2018, **19**(3). doi: 10.17169/fqs-19.3.3099.
58. Zook, M.; Barocas, S.; Boyd, D. *et al.* Ten simple rules for responsible big data research. *PLoS Comput. Biol.,* 2017, **13**(3): e1005399. doi: 10.1371/journal.pcbi.1005399.
59. Cheah, P.Y.; Jatupornpimol, N.; Hanboonkunupakarn, B. *et al.* Challenges arising when seeking broad consent for health research data sharing: A qualitative study of perspectives in Thailand. *BMC Med. Ethics,* 2018, **19**(86). doi: 10.1186/s12910-018-0326-x.
60. Poth, C.N. Rigorous and ethical qualitative data reuse: Potential perils and promising practices. *Int. J. Qual. Methods*, 2019, **18**. doi: 10.1177/1609406919868870.
61. Skovgaard, L.L.; Wadmann, S. & Hoeyer, K. A review of attitudes towards the reuse of health data among people in the European Union: The primacy of purpose and the common good. *Health Policy,* 2019, **123**(6), 564–571. doi: 10.1016/j.healthpol.2019.03.012.

62. Aggarwal, N. & Floridi, L. Ethics of data publication in the context of asylum claims. *Soc. Sci. Res. Network*, 2018.
doi: 10.2139/ssrn.3263377.
63. Curty, R.; Yoon, A.; Jeng, W. & Qin, J. Untangling data sharing and reuse in social sciences. *In* Proceedings of the Association for Information Science and Technology banner., 2016, **53**(1), 1-5.
doi: 10.1002/pra2.2016.14505301025.
64. Sánchez, D.; Martínez, S. & Domingo-Ferrer, J. Comment on 'Unique in the shopping mall: On the reidentifiability of credit card metadata'. *Science*., 2016, **351**(6279), 1274.
doi: 10.1126/science.aad9295.
65. Sugiura, L.; Wiles, R. & Pope, C. Ethical challenges in online research: Public/private perceptions. *Research Ethics*. 2017, **13**(3-4), 184–199.
doi: 10.1177/1747016116650720.
66. Thomas, D.; Pastrana, S.; Hutchings, A.; Clayton, R. & Beresford, A. Ethical issues in research using datasets of illicit origin. *In* Proceedings of IMC '17, London, UK, November 1–3, 2017, 18 pages.
doi: 10.1145/3131365.3131389.
67. Slavnic, Z. Research and data-sharing policy in Sweden – neoliberal courses, forces and discourses*. *Prometheus*, 2017, **35**(4), 249–266.
doi: 10.1080/08109028.2018.1499542
68. Helbig, K. Research data management training for geographers: First impressions. *ISPRS Int. J. Geo-Inf.,* 2016, **5**(4), 40.
doi: 10.3390/ijgi5040040.
69. Kim, Y. & Yoon, A. Scientists' data reuse behaviors: A multilevel analysis. *J. Assoc. Inf. Sci. Technol.*. 2017, **68**(12), 2709–2719.
doi: 10.1002/asi.23892.
70. Hara, Y. & Kuwahara, M. Traffic monitoring immediately after a major natural disaster as revealed by probe data – A case in Ishinomaki after the Great East Japan earthquake. *Transp. Res. Part A: Policy Pract.* 2015, **75**, 1–15.
doi: 10.1016/j.tra.2015.03.002
71. Sandbrook, C. The social implications of using drones for biodiversity conservation. Ambio. 2015, **44**, 636–647.
doi: 10.1007/s13280-015-0714-0.
72. Kujala, R.; Weckström, C.; Darst, R.K.; Mladenović, M.N. & Saramäki, J. A collection of public transport network data sets for 25 cities. *Scientific Data*. 2018. 5.
doi: 10.1038/sdata.2018.89.
73. Custers, B. & Bachlechner, D. Advancing the EU data economy: Conditions for realizing the full potential of data reuse. *Information Polity*. 2017, **22**(4), 291–309.
doi: 10.3233/IP-170419.
74. Bierer, B.E.; Crosas, M. & Pierce, H.H. Data authorship as an incentive to data sharing. *N. Eng. J. Med.*. 2017, **376**, 1684–1687.
doi: 10.1056/NEJMsb1616595.
75. Rolland, B. & Lee, C.P. Beyond trust and reliability: Reusing data in collaborative cancer epidemiology research. in Proceedings of the 2013 Conference on Computer Supported Cooperative Work 435–444 (ACM, 2013).
doi: 10.1145/2441776.2441826.

CONTRIBUTORS

Dr Juan-José Boté is an Assistant Professor and member of Departament de Biblioteconomia, Documentació i Comunicació Audiovisual & Centre de Recerca en Informació, Comunicació i Cultura, Universitat de Barcelona. His areas of interest include digital preservation and cultural heritage. His contribution to the current study included literature collection, literature analysis and discussion of the findings.

Dr Miquel Térmens is a Professor and member of Departament de Biblioteconomia, Documentació i Comunicació Audiovisual & Centre de Recerca en Informació, Comunicació i Cultura, Universitat de Barcelona. His areas of interest include digital preservation and data auditing. His contribution to the current study included discussion of the findings and final supervision.