Ethical Mining – A Case Study on MSR Mining Challenges

Nicolas E. Gold
n.gold@ucl.ac.uk
CREST, Department of Computer Science
University College London
London, United Kingdom

Jens Krinke
j.krinke@ucl.ac.uk
CREST, Department of Computer Science
University College London
London, United Kingdom

ABSTRACT

Research in Mining Software Repositories (MSR) is research involving human subjects, as the repositories usually contain data about developers' interactions with the repositories. Therefore, any research in the area needs to consider the ethics implications of the intended activity before starting. This paper presents a discussion of the ethics implications of MSR research, using the mining challenges from the years 2010 to 2019 as a case study to identify the kinds of data used. It highlights problems that one may encounter in creating such datasets, and discusses ethics challenges that may be encountered when using existing datasets, based on a contemporary research ethics framework. We suggest that the MSR community should increase awareness of ethics issues by openly discussing ethics considerations in published articles.

CCS CONCEPTS

• Software and its engineering; **• Social and professional topics → Codes of ethics**; **• General and reference → Empirical studies**;

KEYWORDS

research ethics, mining software repositories

ACM Reference Format:
Nicolas E. Gold and Jens Krinke. 2020. Ethical Mining – A Case Study on MSR Mining Challenges. In *17th International Conference on Mining Software Repositories (MSR '20), October 5–6, 2020, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3379597.3387462

1 INTRODUCTION

There have been a large number of papers that mine data contained in software repositories. A repository contains considerable information about the author of code as a by-product of their interaction with it, and with their collaborators. In studying this data, the researcher is in effect directly or indirectly studying the person through their data. Oezbek [51] identified that open-source software research (including data mining) involves humans as participants, collaborators, or data sources and thus requires ethics review.

Much research focuses on open-source software repositories and it is widely believed that, as the software is published open-source, no ethics issues will arise (similar to studying published literature). However, there is a difference between the publication of source code (by intentionally applying a licence to the code itself) and the incidental public availability of other data (the repository) that typically lacks such a manifest act of publication. Licences that are FSF or OSI-compliant permit freedom to study [38] or do not restrict the purpose of use [56] and therefore can offer a good ethics defence for studying the code without the need for further ethics review. As repository data is not typically licensed like this, repository-focused studies are therefore far more likely to raise ethics issues than code-focused studies.

Where the potential for ethical issues is identified, compliance with legal frameworks like the GDPR [77] or the The California Consumer Privacy Act of 2018 (CCPA) [21] may be cited as sufficient. Although law and ethics are linked, they are not necessarily the same thing (Hand [43] characterises ethics as guiding what a person *should* do, and the law as what they *must* do). Legal compliance is often considered within ethics (and it is worth noting that there are potential implications of the GDPR for activities that may be considered as profiling developers [79]) but is not necessarily sufficient to achieve ethical safety.

Ethical practice changes in response to societal concerns, technological advances and new ethics theory. Thus practices that were once considered ethical may no longer be so (and vice versa). As ethics theory evolves (for example, data ethics has only emerged as a discrete branch of ethics in recent years [20]) so should ethics and research practice. It is therefore important to periodically look afresh at a research area to consider its current ethical situation (a point also made by Hand about data ethics in general [43]). There are examples of past practices being revisited: email addresses were removed from the GHTorrent data dump in March 2016 [9] and age information was removed from Stack Overflow's public data as part of an audit for the GDPR [22].

It is only recently that MSR researchers have become more aware that repository mining can raise ethics issues (for example, the"Ethical MSR" Discussion Session at the MSR conference 2019 demonstrated both the interest and the potential limits to awareness and training within the community). Indeed as a broader field, Information and Communication Technology Research (ICTR) as a whole is still coming to understand the breadth of ethical issues that it may need to address (see the Menlo Report [28]); Hand describes this as a qualitative change from a focus largely on technical matters [43]. There are some signs that this change is taking place, e.g. Stahl et al. [73] found several papers concerned with Internet-based research in their survey of ethics in computing. They

MSR '20, October 5–6, 2020, Seoul, Republic of Korea
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *17th International Conference on Mining Software Repositories (MSR '20), October 5–6, 2020, Seoul, Republic of Korea*, https://doi.org/10.1145/3379597.3387462.

identified a range of questions that these papers raise including issues of consent, data ownership, method replication, recruitment, respect for privacy, the difficulty of delineating public and private spaces, and data anonymisation. Interestingly some of these issues were also earlier highlighted by Berry [18], suggesting resolution of the ethical matters involved is not straightforward. Research guidance like the Menlo Report [28] was published less than eight years ago, the GDPR came into force in May 2018, and the California Consumer Privacy Act only came into force in January 2020. One must thus be realistic in considering previous work (evaluating it only against the well-understood ethical standards of its time) but also forward-looking in seeking opportunities to improve research practice in future.

In this paper, we explore some ethical issues that may arise in the future study of software repositories in the light of recently-published ethical and legal frameworks, and grounded in the kinds of data that have been used previously for MSR research.

N.B. At the outset, we want to establish clearly that our intention in this paper is to promote and support the development of ever-improving ethical research in future. Thus our arguments and analysis herein are not intended in any way as criticism of other authors, and for that reason we have avoided attempting to retrospectively analyse ethics issues that might have arisen during the course of previous research. Where we have discussed issues that may *in future* arise in respect of the types of data used by MSR challenges, our references to the previous challenge descriptions are again, intended only to ground our identification of the types of data used by the community, not to indicate that such issues should have been discussed at the time those challenges were set (see discussion above about the changing nature of ethics). It is also worth noting that the MSR community is not unique in using these kinds of data and thus the points we raise relate to all such research. As we note elsewhere in the paper, our assumption is that all ethics issues considered relevant at the time of prior research were addressed satisfactorily in the context of the time and individual researchers' institutional and national requirements.

The first step in considering how ethics issues may arise in repository studies is to identify an appropriate ethics framework to apply, e.g. BPS [75, 76], BERA2018 [19], RESPECT [26], the Menlo Report [28], and the Association of Internet Researchers guidelines [47]. To some extent, the choice is affected by the specific nature of the research problem to be addressed (e.g. some studies might be best characterised as digital social science, or internet-mediated social research).

In this paper we have adopted the Menlo Report framework as the lens through which to examine ethics issues. It is specific enough to ICT research to be useful, yet broad enough to encompass a wide range of research questions and activities (for example, see the Companion Report for a range of applications [29]).

As MSR research is very diverse, the MSR mining challenges of the past years are used as a case study to help identify in concrete terms the data used in the field. For each of the underlying data sets, we identify the ethics issues that may need to be considered in creating and/or using them (or similar datasets) in future work. Moreover, we also survey the published papers in the Mining Challenge Track to identify any ethics issues that were previously raised.

The paper is structured as follows. The next section discusses the framework of the Menlo Report and how it is applied to MSR research. Section 3 analyses the data sets of the mining challenges and their specific ethics issues. Section 4 discusses our observations and recommendations. Sections on threats to our research, related work, and conclusions follow.

2 USING THE MENLO REPORT FOR MSR RESEARCH

The Menlo Report [28] proposes a framework for ethics guidelines for Information and Communication Technology Research (ICTR), based on the principles of the 1979 Belmont Report [49]. It is a framework that can be used for any ICT research and addresses ethics issues that typically arise in MSR research.

The Menlo Report sets out four principles for the consideration of ethics, three from the Belmont Report [49] and the fourth added for the evolving nature of the legal contexts of privacy, and information and systems assurance [28]. A summary of the principles and how they are applied is shown below [28]:

(1) *Respect for Persons* (voluntary participation following informed consent; respecting individuals as autonomous and able to determine their own best interests; respect individuals impacted by, but not the targets of, research; protect those with limited autonomy)
(2) *Beneficence* (do not harm; maximise benefit; minimise harm; systematically assess risks of harm/benefit)
(3) *Justice* (consider each individual equally; fairly distribute research benefits considering need, effort, societal contribution and merit; fairly select subjects; allocate burdens equitably)
(4) *Respect for Law and Public Interest* (due diligence; transparency in methods and results; accountability).

The Report proposes a standard method to operationalise the principles: **identification of stakeholders and informed consent; balancing risks and benefits; fairness and equity; and compliance, transparency and accountability**. We now examine these stages in the MSR context. Since informed consent is a significant area, we have separated this from the identification of stakeholders.

2.1 Identification of Stakeholders

The Menlo Report identifies potential stakeholders as: ICT Researchers; Human Subjects, Non-subjects, and ICT Users; Malicious Actors; Network/Platform Owners and Providers; Government (Law Enforcement and Non-Law Enforcement); and Society. For MSR research, the most relevant are likely to be ICT researchers, and human subjects, non-subjects, and users. The range of human subjects may be wide but would typically be those who contribute to a repository. There may also be users who post bug reports and people who are not part of the core team but appear in the repository data. Those impacted by the research but not part of it may include the repository hosting site who may wish to be consulted before agreeing to the use of their users' data for research, and the organisations for which contributors work (e.g. if the research aimed to characterise the relative contributions of developers and observed that one organisation contributed less than others in contradiction to that organisation's marketing).

Malicious actors also cannot be ignored for MSR research. For example, MSR research can uncover security vulnerabilities which could be exploited (requiring to follow responsible disclosure principles), or MSR research on developer behaviour could accidentally uncover malicious behaviour by a developer.

2.2 Informed Consent

Informed consent falls within the Menlo principle of Respect for Persons. In particular, are developers "participating" voluntarily (via their code/repository records) and did they give informed (active) consent? Does the research respect their autonomy? Does the research respect those impacted by it but not part of it (i.e. people other than the authors)? Does the research protect those with limited autonomy?

Informed consent in internet-based research is a complex area in general. Social media research shows that people do not always read or understand terms and conditions and may not understand the public nature of the information they post [17, 74] (in an MSR context this may involve forums, issue trackers, and/or repository commits). The expectation of participants may be at variance with the intended use of the data by researchers. It is likely that those involved in open-source software development have a greater awareness of the public nature of what they are doing than non-specialist users of social media sites, but that cannot be taken as de-facto agreement to research participation: it was not the purpose for which the data was provided. Contributors may be happy to provide code and work with other developers but not happy to have their activity reviewed or commented on by researchers, at least not without having an opportunity to assess for themselves the risks and benefits in advance. This is different to the code-study situation where a developer's application of a licence offsets some of the ethics concerns. To appropriate contributors' activity data in this way could be considered an affront to their autonomy.

Terms and conditions are not always a panacea in this case either as various problems arise: difficulty in determining who to approach for consent [18], the need to actively consent rather than passively agree terms without reading them [43], the assumed adequacy of terms and conditions to cover ethics conditions for studies and the problem of relying on such terms as an ethics defence [17], and finally, the changing participation in open-source developer communities meaning that it may not be possible to seek consent from individuals [51].

2.3 Balancing Risks and Benefits

It is important to establish potential contributions and threats before starting research so that they can be balanced. Research can only have benefits if its results can be trusted and if there is value in them. Current practices address this in a post-hoc fashion: Papers usually highlight their scientific contributions (value) and discuss threats to validity (trust). Moreover, peer review ensures a higher level of trust in the outcomes. Traditionally, threats to validity focus on matters relating the value of the enquiry to things that might confound the results, not the risks to those involved in the enquiry itself. Introducing ethics reasoning adds an additional dimension: risks (threats) to those involved in the research directly and indirectly. Research that tilts too far towards risk may undermine its otherwise valuable contribution because of the harm that could result. For example, if one desired to research a community repository involved in building open-source malware, executing that malware as part of the study might be highly risky to many people.

MSR-type research (in common with other parts of data science) is often done with a mindset of "here is a dataset, let's see what we can find". This is risky because the same data can be used in many ways, some safe, some risky to the participants.

The need for ethical use of repository data (including source code and repositories themselves) is succinctly captured in the ethics policies of the Software Heritage Archive Ethical Charter [71] where to use the archive requires a consideration of the potential harms, the protection of personal data, and taking care of derived data and results. This extends to its policy on mirrored copies [70] and load management on its own server (disallowing massive-scale data extraction in order to equitably serve its users [72]). As a specific example, it also clearly states that mass-mailing of developers using the information in repositories constitutes misuse of the data.

Areas of potential harm in MSR (and software research generally) may relate to observations and judgements of practices, impacting on the individual's reputation, e.g. profiles of contribution rates and success, or code quality. Making such claims about individuals may not only damage their or their organisation's reputation, but may reflect negatively on the researcher making the claims and potentially the researcher's organisation and funding body.

The conclusions drawn could have consequences for a developer's ability to participate in future projects and may affect their ability to secure a job [51]. If they are among the increasingly prevalent group of commercially-situated contributing developers, it could have direct and immediate consequences for their current professional life and reputation. Whilst the data for these conclusions is publicly observable, the potential harm arises from the attention drawn to a particular aspect of that data by the research: the act of research creates the risk. Hosting sites often provide this kind of information themselves in the form of metrics (e.g. GitHub [54]) but it is not known whether the ethics implications have been considered.

2.4 Fairness and Equity

This area concerns the potential for societal contribution from the research, the fair selection of subjects, the availability of results, and the equitable treatment of all developers involved in a study. In general, these areas do not raise MSR-specific concerns. Researchers typically articulate the potential scientific and societal benefits of better understanding the properties of software and methods that operate on it in terms of the significance of their work. Results are published and made as widely available as possible.

Fair selection and equitable treatment is somewhat harder to attain however. In research involving people, fair selection is often addressed using random sampling within a population. That can be harder to achieve for code and repository-related research and there is a risk that certain systems and repositories become frequently used and overall, potential harms become concentrated on them, meaning that additional care may be needed in the selection of repositories for study.

2.5 Compliance, Transparency, and Accountability

It should be noted that the authors are not legally trained and discussion of legal matters in this paper should not be construed as, or used as, legal advice.

In the context of ethics compliance, legal matters cannot be ignored, partly given the fourth principle of the Menlo Report framework that emphasises legal compliance [28] and the fact that legal compliance is part of ethical data handling [20]. The laws in force where research is being undertaken may interact (aligning or sometimes conflicting) with ethics management even though the specific requirements for each may be separate.

For example, in data analysis research with non-anonymous data, the EU's General Data Protection Regulation (GDPR) [77] applies to all countries and researchers within the European Union. The GDPR also contains requirements for organisations outside the EU that process EU citizens' data and it therefore has implications on a potentially global scale (e.g. see the examples cited by Kshetri and Voas [46]). These may affect how researchers acquire and process data, but also how they collaborate with colleagues in other countries. It is worth noting that the GDPR has certain exceptions for scientific research which affect the legality, but which do not affect what is ethical.

"Processing" personal data (a concept that under the GDPR captures a very broad range of activity including acquisition, storage, and analysis through to deletion) requires an appropriate legal basis to be selected to comply with the GDPR. In the context of open-source software, the Linux Foundation GDPR guidance [79] indicates that processing commit data in the context of FLOSS development is likely lawful on the "legitimate interests" basis (although the argument is more nuanced than space reasonably permits here and readers are recommended to read the cited guidance and consult legal counsel where appropriate). Of particular note for researchers using FLOSS data from repositories, the Linux Foundation guidance also notes that "profiling" under the GDPR (e.g. analysing or predicting work performance, reliability, behaviour and so forth) usually requires *explicit consent* from the person concerned, and that confers rights for them to withdraw that consent at any time [79]. This could have implications for publications based on the data since the publications would no longer be able to refer to the data if consent was withdrawn in the future.

Although significant in impact, the GDPR and the CCPA (and other data protection legislation) are not the only legal frameworks that may apply (Broad et al. [20] also note laws around confidential information and anti-discrimination as applicable to data handling). In the US, the HIPAA legislation [52] requires certain safeguards on medical information. Although these areas may not appear to be immediately connected to research on open source software and repositories, the systems being studied may be impacted by them. Intellectual property law (in particular copyright) is also significant in MSR-related research (although patents may also be relevant and as Broad et al. [20] note, database law). Copyright law is the foundation of open-source code licencing. There are many standard licences used for open source development, e.g. see the lists of those compliant with the FSF and OSI conditions for "free" [37] or "open source" [55] software. Being openly available is not enough, there must be explicit provision of licence and it needs to cover both outbound use and inbound contribution either implicitly or through contributor agreements [53]. However, such licences tend only to apply to the code, not the repository metadata that is often used in MSR research. Repository metadata is in effect covered by contract: the agreement of the user of a repository to abide by the terms and conditions of the hosting site and the repository itself.

Compliance in an ethics context is thus a case of working within the legal frameworks and the terms and conditions applicable to the data being used. This might be seen as a necessary but not necessarily sufficient condition to achieve ethics justification for the work. One potential issue to consider is that publicly available datasets may not be free of ethics issues [81] and it is necessary to assess them before use. For example, datasets created before the GDPR came into force may no longer be GDPR compliant.

Matters of transparency and accountability are perhaps easier to resolve in the MSR research context. Researchers normally identify themselves in their outputs and take responsibility for them, making the methods as transparent as possible. Researchers in commercial settings may need to pay more attention to this area as they may be restricted by commercial constraints in terms of what can be reported and how, and thus may need to seek a greater degree of ethics oversight. The balance between the societal and scientific benefit of research and the potential harms to participants may be harder to demonstrate where the primary beneficiary is the organisation itself rather than the scientific community through open dissemination (Benbunan-Fich [17] discusses this with particular reference to online experiments involving deception). Transparency also applies in the conduct of the research itself, and there are therefore challenges to how researchers should identify themselves to the communities whose data they are researching.

3 ETHICS ISSUES IN MSR MINING

As described in the Introduction, to seek some concrete evidence for our hypothesis that ethics issues have not been widely discussed in MSR research, we undertook a simple analysis of papers (and other sources) published in the years 2010 to 2019 of the MSR mining challenges. In a first step, a keyword search was done to identify papers that were discussing ethics issues. We searched for `ethic*` as the primary keyword and used the following secondary terms (concepts related to ethics): `threat`, `anon*`, `privacy`, `confidential*`, and `consent`. The sources we considered for the keyword search included the MSR Mining Challenge website of each year, the website describing the dataset(s) used in the mining challenge, and the papers explaining the mining challenge and/or the underlying datasets. We analysed the results of the keyword search by checking whether every occurrence of a keyword occurs as part of a discussion on ethics issues or related topics.

In a similar way, we also analysed the 102 papers that were published for the Mining Challenge Track in the years 2010 to 2019. We applied the same approach using a keyword search as before and investigated all occurrences of the keywords for discussions about ethics implications. We also did a similar keyword search but this time using the keyword `threats` in order to identify the number of papers discussing threats to validity. As discussed in

Section 2.3, discussions of threats to validity belong to the balancing of risks and benefits.

As might be expected given that collective recognition of the breadth of potential ethics issues is only recent, we found some, but not extensive, discussion: one mining challenge discussed ethics issues in some detail, and a few papers discussed the anonymity of the underlying datasets in the research that built on them. Thirty papers contained discussions of threats to validity (perhaps a section of a paper best suited to the discussion of ethics issues). Using the Menlo Report to unpack potential issues for future research thus seems a useful exercise.

Table 1 shows an overview of the mining challenges. The first column shows the year of the dataset and the second column lists the publications we consulted for the dataset information. The next block of six columns show the type of datasets the challenge used. The last two columns show the number of accepted papers for the Mining Challenge Track in the corresponding year and in how many of the papers we could find a "Threats to Validity" discussion.

Instead of discussing the challenges year by year, we focus on the types of data sources used by the mining challenges. We identified six different data sources for the mining challenges:

IDE Events. One challenge uses a dataset that is not extracted from a software repository: The dataset has been created by capturing events inside an IDE.

Version Control Data. Many challenges use data from version control systems, i.e. data from CVS, Subversion, Git, or Mercurial. The challenges use copies of the repositories or aggregate them into new datasets.

Build Logs. Version control systems often use some kind of Continuous Integration (CI) system to automate building the software. If the build results are archived, they can provide data for research into testing and building practices.

Stack Overflow. Stack Overflow provides official dumps of their data and (subsets of) the dumps have been used as challenges directly, or inside a dataset aggregating historic information.

Issue Tracker Data. Some challenges use data extracted from issue tracker systems like Bugzilla.

Mail Archives One challenge includes mailing list archives. Mailing lists are flexible and can be used for different purposes, e.g. issue tracking, code review, Q&A forums, etc.

We now discuss the different types of datasets listed above and the ethics issues such data may raise in future.

3.1 Mining IDE Events

Capturing IDE event data involves human subjects directly and is thus a classic example of empirical software engineering research. Gathering and using this kind of data raises typical ethics issues of consent and privacy among other things, and indeed these are discussed on the KaVE project website (from which the 2018 challenge data was drawn) and the associated PhD thesis [59] that contains a section on privacy in which anonymisation, profiling, informed consent, incentives for participation, and legal issues are considered.

Identification of Stakeholders and Informed Consent. As with any experiments with human subjects, informed consent to undertake the study and (if desired) future research is necessary. Assuming consent to further use is given, subsequent research using the resulting dataset would not need further informed consent as the data was consented for that use (and it is likely it would have been anonymised).

Table 1: Overview of the Challenges by Year.

Year	Papers	Version Control Data/Metadata	Issue Tracker Data	Mail Archives	Stack Overflow	Travis Logs	IDE Events	Published Papers	Threats to Validity
2010	[44, 45]	×	×	×				6	0
2011	[64, 65]	×	×					5	1
2012	[66, 67]	×	×					6	1
2013	[5]				×			12	1
2014	[13, 41]	×						8	2
2015	[87]				×			14	5
2016	[30–32, 50]	×						10	4
2017	[15, 16, 80]					×		14	2
2018	[58–61]						×	13	7
2019	[7, 8, 10–12, 27]				×			14	7

Balancing Risks and Benefits. The main mechanism to protect participants and their organisation is through anonymisation (including not revealing any industrial partners). If there are going to be industrial or student participants, one needs to consider whether there are additional risks as employees or students (e.g. profiling by or pressure to participate from employers or teachers). Such power relationships apply in other areas too, e.g. if a lead maintainer wished to give permission for their project repository to be studied and pressurised others involved to give permission too.

Fairness and Equity. Fairness and equity considerations seek to ensure that the burden and benefits of research are equally distributed among those involved, and the wider public. This might involve ensuring a mixture of participant types, or drawing on different companies or domains but then publishing to all. For example, the KaVE dataset contains data from a mixture of industrial, research, private, and student participants.

Compliance, Transparency and Accountability. It is important that all involved in a project, particularly where direct observation of work practices are involved, are aware of what is intended and how it complies to policy and legislation. It is therefore important not just to seek informed consent from participants, but also to consider those who may be gatekeepers to the research (e.g. employers, repository admins, platforms). As a good example, the KaVE project's website [58] states that "the captured feedback structure was discussed with the privacy council of a large German IT company and complies to German privacy requirements."

3.2 Mining Version Control Data

Data generated from Version Control Systems like CVS, Subversion, Git, or Mercurial is the typical data mined in MSR-type research, either as a direct copy, or in the form of interaction metadata, and/or in combination with other datasets.

Identification of Stakeholders and Informed Consent. Obtaining consent from users of version control system is usually difficult to impossible (as discussed above). One alternative may be to seek consent from the organisations providing the repositories. Some organisations have general terms and conditions that license the data shared with them, which usually includes any data contained in their repositories. For example, the Eclipse Foundation requires users to agree that any stored or shared information will be subject to a (very much nonrestrictive) CC0 1.0 Creative Commons [23] license. Whilst the data may be licensed, there may still be specific risks raised by particular research that requires informed consent from each potential participant. If consent cannot be practically obtained, one may have to consider seeking a consent waiver from the appropriate IRB/REC in each circumstance and with appropriate justification (see the Menlo Report [28, pp10–11]). This would likely require an argument to be put forward on the basis of the content and clarity of the terms and conditions signed up to by developers, their likely expectations of how their data would be used, and the potential harms in using it without their explicit consent. In particular, it is important for researchers to consider how such an argument would be sustained in future if a developer removed their information from a repository, leaving just the software behind: if the presence of the data “in public” is a key aspect of the consent-waiver argument, once that data is no longer public, will the argument still hold?

Balancing Risks and Benefits. Any research using metadata from version control systems needs to consider the risks to their users. Datasets aggregating metadata create an increased risk to the users as they usually aggregate or link users of different repositories so that they are represented by unique user objects. The aggregation of users over large amounts of data would allow profiling, i.e. the automated processing of personal data to evaluate certain things about an individual.

Anonymisation of such data is almost impossible to achieve as re-identification of the data is almost always possible through the code changes themselves. However, anonymisation and pseudonymisation should still be used to lower the risk for the developers and the researchers. That such ethics concerns are important can be seen in the reports of legal and privacy concerns raised in respect of the GHTorrent dataset leading to email addresses being removed from the data dump in March 2016 [9].

Fairness and Equity. Inequitable selection of subjects is unlikely as the metadata contained in repositories does not usually include sensitive information like gender or religion.

Compliance, Transparency and Accountability. The usage of repositories as provided by organisations is regulated by their terms and conditions. Usually, such terms and conditions are not specifically crafted for the use of the repositories and their metadata for purposes other than as software development. It is therefore necessary to carefully consider the terms and conditions under which the software contained in the repositories is provided and the terms and conditions for the data provided by the organisations' websites. For example, Eclipse repository content is subject to the Eclipse Foundation Software User Agreement [33], but the repositories themselves are subject to the Eclipse.org Terms of Use [34].

3.3 Mining Build Logs

Build logs typically contain detailed data about the result of a build and the commit for which the build was triggered.

Identification of Stakeholders and Informed Consent. While the build logs themselves do not usually contain identifiable information, they are linked to specific commits, which, as discussed above should be considered potentially identifiable information. Therefore, the same considerations for Mining Version Control Data apply here in terms of ethics, and arguments around consent may need to rely on the clarity and comprehensiveness of those terms (for example, Travis CI has a detailed privacy policy [83]).

Balancing Risks and Benefits. The risks in this type of data largely resolve around the difficulty of anonymisation since commit-ids can be resolved to committers and their personal data. Analysing build trends can reveal negative characteristics and when linked to individuals or projects could damage their reputation.

Fairness and Equity. Inequitable selection of subjects is unlikely as the metadata contained in build logs or repositories does not usually include sensitive information like gender or religion.

Compliance, Transparency and Accountability. The creation and use of build log data will not only have to consider compliance, transparency and accountability for accessing and using the logs, but also for the repositories for which the builds have been created.

3.4 Mining Stack Overflow

Stack Overflow is the go-to Q&A website for programmers. Stack Exchange (the organisation behind) provides official dumps of the Stack Overflow data. Using the Stack Overflow data in research raises similar ethics considerations to research in other areas using secondary data from websites.

Identification of Stakeholders and Informed Consent. When users register with Stack Overflow, they are referred to the Terms of Service, which explicitly states that by registering one agrees to make all content available under CC-BY-SA Creative Commons license terms, including regular dumps of the content, now called “Creative Commons Data Dump”. Consent to the creation and sharing of the dataset has therefore been given, but this does not necessarily imply that informed consent has been given to any and all research using the dataset as there may be particular risks that require explicit consent. The clear and explicit licence does give strong support to an ethical defence for data use in general.

Balancing Risks and Benefits. While the creators of the Stack Overflow dump ensured to not accidentally release personally identifying information [4], it is up to the specific researcher using Stack Overflow data to balance risks and benefits. The overall benefit of using the Stack Overflow data for research is evidenced by thousands of research papers based on it. However, there are clear risks to the users posting content on Stack Overflow: While users have the option to not reveal personal data, many users opt to include personal data like their real name, their website, their location, or their GitHub username in their public profile which makes them identifiable. In particular research that aims at observing and analysing user behaviour of Stack Overflow participants needs to protect users from the risks of revealing behaviour that could negatively affect them (personally or professionally).

One potential ethics consideration should be that the group of Stack Overflow users contains minors and research with minors participating usually requires additional ethics procedures (including additional consent arrangements).

Fairness and Equity. Stack Overflow does not reveal data about gender and race but does contain location data which should not be used to arbitrarily target persons or groups (and depending on the nature of processing, may require additional GDPR-related considerations). Until 2018 the data dump contained the age of the user, which is considered sensitive information and was removed from public data as part of an audit for GDPR [22].

Compliance, Transparency and Accountability. The CC-BY-SA Creative Commons license terms under which Stack Overflow data is made available are favourable to researchers as it makes it easy to comply with them. However, there is a still a risk to violate licenses as Stack Overflow is known to contain code fragments that potentially violate their original license [1, 62].

3.5 Mining Issue Trackers

Data from issue trackers come in various forms, for example, they are aggregated in the Ultimate Debian Database [25] or dumps of issue tracker systems like in Eclipse and Netbeans. The Eclipse Foundation and the Netbeans Foundation have previously provided dumps of their issue trackers with personal information such as email addresses or users' real names removed. Others (Chrome, Firefox) have previously declined to provide a dump of their issue trackers to avoid making security bugs public.

Identification of Stakeholders and Informed Consent. Issue tracker data is usually covered on the terms and conditions or privacy policies. For example, Debian's Privacy Policy [24] explicitly mentions their bug tracker and states that any information, including names and email addresses as part of email headers will be archived and publicly available. Thus once again, it may be possible to argue to a REC or IRB for waiving the usual informed consent requirements on the basis of the terms and conditions.

Balancing Risks and Benefits. Issue tracker information could be used for profiling users, therefore putting them at risk.

Fairness and Equity. Inequitable selection of subjects is unlikely as the data contained in issue trackers does not usually include sensitive information like gender or religion.

Compliance, Transparency and Accountability. One issue to consider during the analysis of issue tracker data is the disclosure of discovered vulnerabilities. One can usually assume that the organisation behind the issue tracker or the user reporting the vulnerability have followed responsible disclosure procedures. It is different when only the research analysing the tracker data is identifying that a reported issue is actually a vulnerability. The proper responsible disclosure procedures need to be followed in such a case.

3.6 Mining Mailing Lists

Mailing list archives capture the communication between developers and as such contain personal information including email addresses.

Identification of Stakeholders and Informed Consent. The FreeBSD's Privacy Policy [78] explicitly mentions their mailing lists and states that "*Information submitted in those reports and lists, including your Personally Identifiable Information, is considered public and will be accessible to anyone on the web. ... The FreeBSD Foundation has no control over the use of that information, including your Personally Identifiable Information.*" Similar to the discussion before, this could be interpreted as consent to collecting the mailing list archives into a dataset. However, users would not necessarily expect that their mail is used for empirical research and, therefore, research using such data needs to consider whether informed consent needs to be acquired, particularly given the completely free and unstructured nature of email communication.

Balancing Risks and Benefits. The inclusion of personally identifiable information comes with the usual risks. In particular, the unsanitised full email addresses allow intentional or unintentional exploitation of the subjects. Beyond this, users may intentionally or unintentionally disclose other aspects of their views or practices.

Fairness and Equity. Mailing list archives do not usually contain sensitive information like gender or religion explicitly. However, there is a risk that the mailing list archives will capture more 'social' discussions which may reveal gender, religion, political interests, etc. and may therefore fall both within the realms of ethics consideration and GDPR special category data.

Compliance, Transparency and Accountability. As mailing lists are used for various purposes, one has to consider all issues raised above for the other dataset types. Mailing lists can serve as issue trackers or as Q&A forums (see the discussion on Stack Overflow), and they can even contain the build logs.

3.7 Combining Datasets

The discussion above was focussed on the discussion of ethics considerations for specific dataset sources. However, often datasets are combined into larger datasets.

When combining datasets, ethics considerations apply to each dataset separately, but also again for the combined dataset. Is informed consent necessary for the combination? Do additional risks occur by combining the datasets? Are the licenses and terms of the combined datasets compatible? The combination may change the risk to individuals (e.g. anonymised datasets in combination can lead to re-identification of individuals although if that can happen, the individual datasets may not be considered anonymous in the first place).

4 DISCUSSION

4.1 Challenges

Researchers using third-party provided data face a dilemma: Should they trust that the data has been collected for the purpose of their intended use in an ethical way, or do they need to try and verify this for themselves? Not trusting that the data has been collected in an ethical way would prevent its usage, yet completely trusting that it has been ethically collected may not meet the ethical duties that a researcher must fulfil (in general, or in respect of their institution's policies). Even assuming that the data has been collected in an ethical way does not mean that the intended research using the provided data does not need to also consider and balance relevant ethics issues. On the contrary, even the use of data that has been collected in an ethical way, including obtaining consent for the use in the challenge or other research requires considerations of ethics issues once again. It is important to distinguish between the data collection and the data usage, as most ethical issues will arise at the usage level. For example, the creation of the CROP dataset [57], a collection of code review data, did not require full review through our institution's Research Ethics Committee, but using it for profiling reviewers' productivity would certainly require full review.

Ethical safety is promoted through a consideration (and possibly review) of: the sources of data, the collection methods used, the intended use to which the data will be put, and the way in which it will be presented. Where ethics issues are identified, it is helpful for them to be discussed in papers in order that research communities can identify and improve good practice, provide opportunities for new researchers to learn and understand the norms and expectations of the community as they are understood at that time, and provide transparency about what has been considered. Many conferences and journals are beginning to require an ethics statement as a matter of course when a paper is submitted. Whether such statements should be separate to the main body of a paper or form part of the Threats to Validity is an open issue. We speculate that in the MSR context (or others), page limits may act as a disincentive to including such sections in the face of reporting the primary results of research.

In times when papers are asked to provide their data to allow replication, there is also a significant challenge to the MSR community to find ways to not expose identifiable data in the research artefacts or replication packages. Perhaps a new approach to replicability is needed where systems or repositories are characterised in terms of the properties that make methods replicable, rather than simply offering replication on the same systems. This would be analogous to medical research where new treatments are demonstrated using replication studies across different sets of patients rather than replicating on the same set each time.

Given the widespread ethics issues in MSR research, one has to assume that review through Research Ethics Committees or Boards will often be necessary. The process of preparing an application for review and getting approval from the Research Ethics Committee or Board is time consuming. Given the limited time between the publication of the mining challenge and the submission deadline, one may speculate about many intended submissions could not gain approval with sufficient time left for actually doing the research.

4.2 Potential Solutions

Solutions to reputational risks could lie in maintaining developer privacy through anonymity: treating systems as the personal data of their authors and applying the kinds of techniques required in human participant research to protect identity. The difficulty is that the effect of linking multiple extant data sources means that even if directly identifying information like names is removed, other content can be used to resolve identity [43]. In the social media context, this would be the content of posts; for code research, it could be the code itself (or quotations or graphs [51]). Thus protecting a developer's identity might require protecting the identity of the systems to which they contributed (so using code excerpts in publications may need to be avoided), creating a tension between the principles of transparency and the protection of participants.

Obtaining consent in the MSR scenario may be difficult because of the etiquette governing the use of repository information. In theory, consent might be sought using the repository records of email addresses on commits or elsewhere, and sending mass email to seek participation. This is ethically problematic for a number of reasons, not least because the email addresses were not supplied for that purpose and to use them without permission in this way would not respect the individuals concerned (the Software Heritage Archive identifies this activity specifically as data misuse [71] and developers are getting annoyed when receiving such emails too often [9]).

Another possibility might include the development of a licence that developers could attach to their profile governing the use of their repository data. Alternatively, the Menlo Report suggests it might make sense to argue for a consent waiver from an oversight body on the grounds of impracticality [28].

Alternative approaches to consent and ethics matters may be found in other areas of internet-mediated research. Tuikka et al. [84] present a recent survey of netnographic research, a method for studying computer-mediated cultures and communities, based on traditional ethnographic methods. As they make clear, ethics questions and practice are still emerging and there is not yet consensus about what approaches (e.g. to consent, identification, confidentiality, and quotation) may be considered ethically just. Townsend and Wallace [82] define "social media" in the context of their ethics guidance to mean any social online data except email. The discussions in issue/bug trackers (and other related fora) would seem to fall within that definition and therefore guidance from the field of social media research may be relevant (although as Townsend and Wallace [82] note, each context is unique and thus researchers and their oversight body have the responsibility to determine appropriate ethical approaches in response to the challenges posed). Sugiura et al. [74] survey a range of ethics frameworks and literature relating to research practices online, reporting experiences of undertaking internet forum-based research, particularly the difficulty of obtaining informed consent in such a context.

The fact that this debate remains open would suggest that researchers studying repositories (and associated data like issue lists and discussions) will need to consider a wide range of methodological and disciplinary approaches to their work, justifying these in some depth when working with their oversight bodies.

In the future, it seems wise for the MSR community to not only consider the ethics implications of their datasets and their research, but openly discuss them. While it is common to discuss threats to validity in detail in papers, one should consider to also discuss "Ethics Considerations" in which ethics issues and risks are presented. For example, the Empirical Software Engineering journal already has a policy that authors should include a section on "Compliance with Ethical Standard" [36]. Moreover, the current page limit for mining challenge papers incentivises authors to not discuss ethics considerations – an incentive to discuss ethics considerations and raise awareness would be allowing such a discussion outside the page limit. Or with the words of Miller and Rosenstein [48]: "A slightly longer article should be a price worth paying for enhanced accountability."

Moreover, future authors of dataset papers could help future users of those datasets by providing a detailed discussion of ethics considerations in the collection of data and its potential applications in research.

4.3 Reflection

In the spirit of following our own advice, we now discuss the ethics issues we considered in relation to this study. Our assumptions and underlying principles are discussed in the Introduction. Since all the work we studied was in the published literature, analysis of the methods presented in the papers is within the legitimate norms of scientific methodological critique (and falls outside our institution's requirement for ethics review). Nonetheless, one must consider that there may be reputational risks to the authors of that work if the conclusions are handled carelessly (as indeed in any discussion of others' published output). We aimed to manage these risks by focusing the investigation on a keyword-based analysis of the papers themselves (with a degree of subsequent manual checking to ensure keywords were correctly identifying what we sought), thus attempting to avoid the imputation of ethics consideration (or lack of consideration) to the authors. The investigation is objective and about the published works (not the authors): it is therefore an investigation of the frequency of *discussion* of ethics issues in the published literature.

The main risk of the analysis would be the discovery of "unethical" behaviour. Since ethics consideration is usually a process of finding balance between benefit and risk, finding truly unethical behaviour was unlikely since we assume that the potential research outcomes were weighed against such risks and the necessary discussions and processes to gain approval for the intended research followed. Thus an absence of discussion does not equate to an absence of ethics consideration: it is simply an absence of discussion.

Whilst one would not normally comment on changes made in readiness for a camera-ready manuscript, in the context of a discussion of ethical practices and in the interests of full transparency, it may be of interest to note that it was helpfully indicated to us during the review process that our original manuscript could be read as blaming previous authors for a failure to consider ethics, something that was never our intention and that we had tried (but clearly not succeeded) to avoid in our writing. This demonstrates that even ethically-aware research can be difficult to get right, that peer review is an important part of encouraging good ethical practice in the research community, and that presentational matters are of equal importance to those concerned with undertaking research itself. We are grateful for the helpful points made in the review process, have taken steps to prepare the camera-ready manuscript with this issue particularly in mind, and hope that we have now succeeded in positioning our work as we originally intended.

5 THREATS

5.1 Threats to Validity

As our investigation of the extent of published ethics discussion only considered datasets that have been used in the mining challenges from 2010 to 2019, our analysis and discussion cannot necessarily be generalised to all datasets, or MSR research using other datasets. Each type of dataset requires its own specific ethics considerations. However, we observed recurring patterns that can be considered for other datasets or research in MSR.

Our observation of the absence of discussion of ethics considerations is based on what has been reported in the papers or other resources discussing the datasets. It is possible that papers or other resources containing discussions of ethics considerations have been missed, or that our keyword range was insufficiently broad to identify them. Moreover, the lack of discussion cannot be used as evidence for a lack of ethics considerations. Instead, it could perhaps suggest that either there were no relevant issues to consider at the time, or that the lack of space led to the omission of the discussion of ethics considerations that occurred in the process of the presented research.

Any discussion on website content or terms and conditions etc. is based on the versions as of December 2019. At the time of the mining challenges, the website content or the terms and conditions may have been different as people and organisations became more aware of ethics considerations. Moreover, introduction of new regulations like the GDPR or the CCPA will have caused changes to website content and terms and conditions. Our discussion of ethics issues therefore may have been different if the website content and terms and conditions were used as of the date of the mining challenges.

5.2 Ethics Considerations

As described in the Discussion Section, before starting the above presented research, its ethics implications were considered in detail. Similar to literature reviews, only published literature has been used for research and at no point were the datasets themselves (or any other dataset) used. Risks to the original authors of the studied papers have been identified and considered. One risk is that our study would identify ethics issues that would have *prima facie* suggested that the previously-published work was unethical (or even unlawful). Given that the studied papers have all been peer-reviewed, this risk has been considered to be very low since the peer-review process reflects the current state of ethical thinking at the time of publication. Another risk is that our study would identify ethics issues that the original authors have not considered and discussing them could have adverse effects on the original authors. We assume that the original authors have considered all the relevant ethics issues applicable at the time to their work and received approval (if necessary according to their local policies) from their resp. Research Ethics Committees or Institutional Review Boards. In addition to the above, we avoided attempting to retrospectively analyse prior work for ethics issues since to use a contemporary lens to view the legitimacy of past work would be inappropriate. As the risk to original authors is low, the legitimate interest in the extent of discussion of ethics considerations in MSR research outweighs this risk.

6 RELATED WORK

Professional codes of ethics (e.g. IEEE-CS/ACM [40], BCS [14], ACM [3]) do not typically address research ethics directly (although the ACM code [3] does so in its illustrative examples). Hand notes that there is no single profession that has responsibility for data science and thus multiple ethics codes may be relevant and contribute in different ways [43]. Ethics has long been an integral part of research in most disciplines, including computing (see Stahl et al. [73]) and software engineering (human studies in particular, see Hall and Flynn [42], Singer and Vinson [69] and Vinson and Singer [86]). However, it can often be seen as relevant only to studies involving face to face contact with people through observation, interview, and survey, and researchers in ICT do not always realise they are engaged in research that falls within the remit of ethics review [28].

At the same time as the 2001 code [40] was developed, the focus turned to ethics issues in empirical software engineering, summarised by a special issue on research ethics for empirical software engineering [68]. In the same issue, Hall and Flynn [42] present survey results collected from 44 Computer Science Departments in UK Universities and highlight a number of issues.

Singer and Vinson [69] discuss ethics issues in empirical studies in software engineering. They reviewed existing codes and abstracted four principles: informed consent, scientific value, beneficence, and confidentiality. These four principles can be, more or less, mapped to the four principles of the Menlo Report and their operationalisation. Most of the presented examples are studies employing human subjects. However, they also discuss issues that arise when analysing source code, which they also discussed in an earlier paper [85]. Vinson and Singer [86] extend their earlier work [69] into a practical guide to ethical research involving humans in software engineering. They discuss ethics issues around the principles of informed consent, scientific value, beneficence, and confidentiality in detail. Moreover, they discuss how to plan for ethics and prepare for review through an Ethics Review Board (i.e. Research Ethics Committee or Board).

This paper only addressed ethics issues that need to be considered when mining open-source software repositories. However, additional issues arise when such research is done on a collaborating company's data. Andrews and Pradhan [2] discuss ethics issues in such contexts.

El-Emam [35] also raises a series of questions about the ethics implications of analysing open-source software, namely informed consent, minimization of harm and confidentiality. German [39] discusses the analysis of CVS repositories and raises multiple questions about ethics issues in such research, but does not answer them.

Robles et al. [63] present the results of an online survey over 2000 open-source contributors. They also present a case study on linking the survey data with data from software repositories. In this context they discuss how sharing and combining data can lead to ethics and legal issues. They discuss the limits of, and approaches to, anonymisation.

Mining software repositories usually does not require researchers to recruit humans for empirical research. However, sometimes it is necessary to validate results from mining repositories with the resp. developers. Baltes and Diehl [9] discuss ethics issues that arise when contacting developers. They highlight the issue that developers on GitHub are contacted too often and may get annoyed. Moreover, they discuss that email addresses were removed from the GHTorrent data dump in March 2016 due to legal and privacy concerns.

The survey of Badampudi [6] is closely related to our work. The authors have surveyed seven articles that would require informed consent which appeared 2016/17 in the Empirical Software Engineering Journal. Despite the journal's policy to require a discussion on ethics issues, only two of the seven surveyed articles contained such a discussion.

7 CONCLUSIONS

Software repositories always contain personal information or identifiers that can be mapped to individuals. Given that repositories are usually publicly available, even supposedly anonymised datasets usually contain sufficient information to allow mapping of the anonymised data to individual developers. Therefore, one usually has to assume that research using an MSR dataset can affect human subjects, requiring careful consideration of ethics implications. It is an often occurring misunderstanding that analysis of publicly available data is free from the requirement of ethics consideration. Particular problems in MSR research are the considerations for informed consent, risks to the subjects, and compliance.

We presented an exposition of the ethics issues that could arise in MSR research drawing on a contemporary ICT research ethics framework: the Menlo Report. We identify typical problems and discuss their implications. Anonymisation of repository data is almost impossible to achieve as re-identification of the data is almost always possible through the code changes themselves. However, anonymisation and pseudonymisation should still be used to lower the risk for the developers and the researchers.

In the future, the MSR community should not only continue to consider the ethics implications of their datasets and their research, but openly discuss them. While it is common to discuss threats to validity in detail in papers, one should consider to also discuss "Ethics Consideration" in which ethics issues and risks are presented. One way to achieve this would be to adopt the policy of the Empirical Software Engineering Journal that authors should include a section on "Compliance with Ethical Standard" [36]. Moreover, future authors of datasets and dataset papers could help future users of those datasets by providing a detailed discussion of ethics considerations in the collection of data and its potential applications in research.

The current page limit for mining challenge papers perhaps incentivises authors to not discuss ethics considerations – an incentive to discuss ethics considerations and raise awareness would be allowing such a discussion outside the page limit.

ACKNOWLEDGMENTS

The authors are very grateful to their colleagues with whom they have had many helpful discussions about these and other ethics issues, in particular: David Binkley, Evangelos Himonides, Simon Parkin, Ross Purves, and Caroline Wardle.

REFERENCES

[1] Le An, Ons Mlouki, Foutse Khomh, and Giulian o Antoniol. 2017. Stack Overflow: A Code Laundering Platform?. In *SANER '17*. 283–293.

[2] Anneliese Amschler Andrews and Arundeep S. Pradhan. 2001. Ethical issues in empirical software engineering: The limits of policy. *Empirical Software Engineering* 6, 2 (2001), 105–110.

[3] Association for Computing Machinery, Inc. (ACM). 2018. *ACM Code of Ethics and Professional Conduct.* https://www.acm.org/binaries/content/assets/about/acm-code-of-ethics-booklet.pdf (last accessed 3rd November 2019).

[4] Jeff Atwood. 2009. *Stack Overflow Creative Commons Data Dump.* https://stackoverflow.blog/2009/06/04/stack-overflow-creative-commons-data-dump/

[5] Alberto Bacchelli. 2013. *Mining Challenge 2013: Stack Overflow.* http://2013.msrconf.org/challenge.php

[6] Deepika Badampudi. 2017. Reporting Ethics Considerations in Software Engineering Publications. In *Proceedings of the 2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM).* 205–210.

[7] Sebastian Baltes. 2019. *Software Developers' Work Habits and Expertise.* Ph.D. Dissertation. Universität Trier.

[8] Sebastian Baltes. 2020. *The SOTorrent Dataset.* https://empirical-software.engineering/projects/sotorrent/

[9] Sebastian Baltes and Stephan Diehl. 2016. Worse Than Spam. In *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM).* arXiv:1707.00838

[10] Sebastian Baltes and Stephan Diehl. 2019. Usage and attribution of Stack Overflow code snippets in GitHub projects. *Empirical Software Engineering* 24, 3 (June 2019), 1259–1295. arXiv:1802.02938

[11] Sebastian Baltes, Lorik Dumani, Christoph Treude, and Stephan Diehl. 2018. SOTorrent. In *Proceedings of the 15th International Conference on Mining Software Repositories (MSR).* 319–330. arXiv:1803.07311

[12] Sebastian Baltes, Christoph Treude, and Stephan Diehl. 2019. SOTorrent: Studying the Origin, Evolution, and Usage of Stack Overflow Code Snippets. In *IEEE/ACM 16th International Conference on Mining Software Repositories (MSR).* 191–194. arXiv:1809.02814

[13] Olga Baysal. 2014. *Mining Challenge.* http://2014.msrconf.org/challenge.php

[14] BCS: The Chartered Institute for IT. 2015. Code of Conduct for BCS Members. https://cdn.bcs.org/bcs-org-media/2211/bcs-code-of-conduct.pdf (last accessed 3rd November 2019).

[15] Moritz Beller, Georgios Gousios, and Andy Zaidman. [n.d.]. TravisTorrent: Synthesizing Travis CI and GitHub for Full-Stack Research on Continuous Integration. In *IEEE/ACM 14th International Conference on Mining Software Repositories (MSR).* 447–450.

[16] Moritz Beller, Georgios Gousios, and Andy Zaidman. 2017. *Mining Challenge.* http://2017.msrconf.org/#/challenge

[17] Raquel Benbunan-Fich. 2017. The ethics of online research with unsuspecting users: From A/B testing to C/D experimentation. *Research Ethics* 13, 3-4 (2017), 200–218.

[18] David M. Berry. 2004. Internet research: privacy, ethics and alienation: an open source approach. *Internet Research* 14, 4 (2004), 323–332.

[19] British Educational Research Association. 2018. Ethical Guidelines for Educational Research, 4th edition. https://www.bera.ac.uk/publication/ethical-guidelines-for-educational-research-2018 (last accessed 3rd November 2019).

[20] Ellen Broad, Amanda Smith, and Peter Wells. 2017. *Helping organisations navigate ethical concerns in their data practices (white paper).* Open Data Institute. https://www.scribd.com/document/358778144/ODI-Ethical-Data-Handling-2017-09-13 (last accessed 2nd November 2019).

[21] California State Legislature 2018. Assembly Bill No. 375 – California Consumer Privacy Act.

[22] Nick Craver. 2018. *Invalid column name 'Age' in Stack Exchange Data Explorer (Answer).* https://meta.stackoverflow.com/questions/368976/invalid-column-name-age-in-stack-exchange-data-explorer#369002

[23] Creative Commons [n.d.]. *CC0 1.0 Universa.* https://creativecommons.org/publicdomain/zero/1.0/legalcode

[24] Debian [n.d.]. *Privacy Policy.* https://www.debian.org/legal/privacy

[25] Debian [n.d.]. *Ultimate Debian Database.* https://wiki.debian.org/UltimateDebianDatabase/

[26] Sally Dench, Ron Iphofen, and Ursula Huws. 2004. *An EU Code of Ethics for Socio-Economic Research.* The Institute for Employment Studies, UK. http://www.respectproject.org/ethics/412ethics.pdf.

[27] Stephan Diehl, Sebastian Baltes, and Christoph Treude. 2019. *Mining Challenge.* https://2019.msrconf.org/track/msr-2019-Mining-Challenge?track=MSR%20%20Mining%20Challenge

[28] David Dittrich and Erin Kenneally. 2012. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803_1.pdf (last accessed 29th October 2019).

[29] David Dittrich and Erin Kenneally. 2013. Applying Ethical Principles Guiding Information and Communication Technology Research: A Companion to the Menlo Report. https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCOMPANION-20120103-r731_1.pdf (last accessed 29th October 2019).

[30] Robert Dyer. 2013. *Bringing ultra-large-scale software repository mining to the masses with Boa.* Ph.D. Dissertation. Iowa State University.

[31] Robert Dyer, Hoan Anh Nguyen, Hridesh Rajan, and Tien N. Nguyen. 2013. Boa: A language and infrastructure for analyzing ultra-large-scale software repositories. In *35th International Conference on Software Engineering (ICSE).* 422–431.

[32] Robert Dyer, Hoan Anh Nguyen, Hridesh Rajan, and Tien N. Nguyen. 2015. Boa: Ultra-Large-Scale Software Repository and Source-Code Mining. *ACM Transactions on Software Engineering and Methodology* 25, 1 (Dec. 2015).

[33] Eclipse Foundation 2017. *Eclipse Foundation Software User Agreement.* https://www.eclipse.org/legal/epl/notice.php

[34] Eclipse Foundation 2019. *Eclipse.org Terms of Use.* https://www.eclipse.org/legal/termsofuse.php

[35] Khaled El-Emam. 2001. Ethics and open source. *Empirical Software Engineering* 6, 4 (2001), 291–292.

[36] Empirical Software Engineering [n.d.]. *Compliance with Ethical Standards.* https://www.springer.com/journal/10664/submission-guidelines#Instruction%20for%20Authors_Compliance%20with%20Ethical%20Standards

[37] Free Software Foundation. [n.d.]. Various Licenses and Comments about Them. https://www.gnu.org/licenses/license-list.en.html (last accessed 3rd November 2019).

[38] Free Software Foundation. [n.d.]. What is free software? https://www.gnu.org/philosophy/free-sw.html (last accessed 28th October 2019).

[39] D.M. German. 2004. Mining CVS repositories, the softChange experience. In *International Workshop on Mining Software Repositories (MSR).* 17–21.

[40] Don Gotterbarn. 2001. Software engineering code of ethics and professional practice. *Science and Engineering Ethics* 7, 2 (June 2001), 231–238.

[41] Georgios Gousios. 2013. The GHTorent dataset and tool suite. In *10th Working Conference on Mining Software Repositories (MSR).* 233–236.

[42] Tracy Hall and Valerie Flynn. 2001. Ethical Issues in Software Engineering Research: A Survey of Current Practice. *Empirical Software Engineering* 6, 4 (Dec. 2001), 305–317.

[43] David J. Hand. 2018. Aspects of Data Ethics in a Changing World: Where Are We Now? *Big Data* 6, 3 (2018).

[44] Abram Hindle. 2010. *MSR Mining Challenge 2010.* http://2010.msrconf.org/challenge/

[45] Abram Hindle, Israel Herraiz, Emad Shihab, and Zhen Ming Jiang. 2010. Mining Challenge 2010: FreeBSD, GNOME Desktop and Debian/Ubuntu. In *7th IEEE Working Conference on Mining Software Repositories (MSR).* 82–85.

[46] Nir Kshetri and Jeffrey Voas. 2020. Thoughts on General Data Protection Regulation and Online Human Surveillance. *Computer* 53, 1 (Jan. 2020), 86–90.

[47] Annette Markham and Elizabeth Buchanan. 2012. Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0). https://aoir.org/reports/ethics2.pdf (last accessed 3rd November 2019).

[48] Franklin G. Miller and Donald L. Rosenstein. 2002. Reporting of ethical issues in publications of medical research. *The Lancet* 360, 9342 (Oct. 2002), 1326–1328.

[49] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research.* https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html (last accessed 2nd November 2019).

[50] Hoan Nguyen and Robert Dyer. 2016. *Mining Challenge.* http://2016.msrconf.org/#/challenge

[51] Christopher Oezbek. 2008. Research Ethics for Studying Open Source Projects. In *Proceedings of 4th Research Room @ FOSDEM.* http://www.inf.fu-berlin.de/~oezbek/pub/OezbekC08_ResearchEthicsForOSS.pdf (last accessed 3rd November 2019).

[52] U.S. Department of Health and Human Services. 1996. *Health Insurance Portability and Accountability Act of 1996 P.L. No. 104-191.*

[53] Open Source Guides. [n.d.]. The Legal Side of Open Source. https://opensource.guide/legal/ (last accessed 2nd November 2019).

[54] Open Source Guides. [n.d.]. Open Source Metrics. https://opensource.guide/metrics/ (last accessed 3rd November 2019).

[55] Open Source Initiative. [n.d.]. Licenses by Name. https://opensource.org/licenses/alphabetical (last accessed 3rd November 2019).

[56] Open Source Initiative. [n.d.]. The Open Source Definition (Annotated). https://opensource.org/osd-annotated (last accessed 28th October 2019).

[57] Matheus Paixao, Jens Krinke, Donggyun Han, and Mark Harman. 2018. CROP: Linking Code Reviews to Source Code Changes. In *Proceedings of the 15th International Conference on Mining Software Repositories.* 46–49.

[58] Sebastian Proksch. [n.d.]. *KaVE Project.* http://www.kave.cc/

[59] Sebastian Proksch. 2017. *Enriched Event Streams: A General Platform For Empirical Studies On In-IDE Activities Of Software Developers.* Ph.D. Dissertation. Technische Universität Darmstadt.

[60] Sebastian Proksch, Sven Amann, and Sarah Nadi. 2018. Enriched event streams. In *Proceedings of the 15th International Conference on Mining Software Repositories (MSR)*. 62–65.

[61] Sebastian Proksch, Sven Amann, and Sarah Nadi. 2018. *Mining Challenge.* https://2018.msrconf.org/track/msr-2018-Mining-Challenge

[62] C. Ragkhitwetsagul, J. Krinke, M. Paixao, G. Bianco, and R. Oliveto. 2019. Toxic Code Snippets on Stack Overflow. *IEEE Transactions on Software Engineering* (2019).

[63] Gregorio Robles, Laura Arjona Reina, Alexander Serebrenik, Bogdan Vasilescu, and Jesús M. González-Barahona. 2014. FLOSS 2013: a survey dataset about free software contributors: challenges for curating, sharing, and combining. In *Proceedings of the 11th Working Conference on Mining Software Repositories (MSR)*. 396–399.

[64] Adrian Schröter. 2011. MSR Challenge 2011. In *Proceeding of the 8th working conference on Mining software repositories (MSR)*.

[65] Adrian Schröter. 2011. *Mining Challenge.* http://2011.msrconf.org/msr-challenge.html

[66] Emad Shihab. 2012. *Mining Challenge.* http://2012.msrconf.org/challenge.php

[67] Emad Shihab, Yasutaka Kamei, and Pamela Bhattacharya. 2012. Mining challenge 2012: The Android platform. In *9th IEEE Working Conference on Mining Software Repositories (MSR)*. 112–115.

[68] Janice Singer and Norman Vinson. 2001. Why and how research ethics matters to you. Yes, You! *Empirical Software Engineering* 6, 4 (2001), 287–290.

[69] Janice Singer and Norman G. Vinson. 2002. Ethical Issues in Empirical Studies of Software Engineering. *IEEE Trans. Softw. Eng.* 28, 12 (Dec. 2002), 1171–1180.

[70] Software Heritage Archive. [n.d.]. Software Heritage: Ethical Charter for Mirrors. https://www.softwareheritage.org/legal/mirrors-ethical-charter/ (last accessed 3rd November 2019).

[71] Software Heritage Archive. [n.d.]. Software Heritage: Ethical Charter for using the archive data. https://www.softwareheritage.org/legal/users-ethical-charter/ (last accessed 3rd November 2019).

[72] Software Heritage Archive. [n.d.]. Software Heritage: Terms of use for bulk access. https://www.softwareheritage.org/legal/bulk-access-terms-of-use/ (last accessed 3rd November 2019).

[73] Bernd Carsten Stahl, Job Timmermans, and Brent Daniel Mittelstadt. 2016. The Ethics of Computing: A Survey of the Computing-Oriented Literature. *Comput. Surveys* 48, 4 (2 2016).

[74] Lisa Sugiura, Rosemary Wiles, and Catherine Pope. 2017. Ethical challenges in online research: Public/private perceptions. *Research Ethics* 13, 3-4 (2017), 184–199.

[75] The British Psychological Society. 2014. *Code of Human Research Ethics.*

[76] The British Psychological Society. 2018. *Code of Ethics and Conduct.*

[77] The European Parliament and the Council of the European Union. 2016. General Data Protection Regulation (EU) 2016/679. *Official Journal of the European Union* (2016). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679 (last accessed 3rd November 2019).

[78] The FreeBSD Project 2012. *FreeBSD's Privacy Policy.* https://www.freebsd.org/privacy.html

[79] The Linux Foundation. 2018. Summary of GDPR Concepts for Free and Open Source Software Projects. https://www.linuxfoundation.org/wp-content/uploads/2018/05/lf_gdpr_052418.pdf (last accessed 3rd November 2019).

[80] The TestRoots Team 2020. *TravisTorrent: Free and Open Travis Analytics for Everyone.* https://travistorrent.testroots.org

[81] Daniel R. Thomas, Sergio Pastrana, Alice Hutchings, Richard Clayton, and Alastair R. Beresford. 2017. Ethical issues in research using datasets of illicit origin. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC)*.

[82] Leanne Townsend and Claire Wallace. 2016. Social Media Research: A Guide to Ethics. https://www.gla.ac.uk/media/Media_487729_smxx.pdf (last accessed 3rd November 2019).

[83] Travis CI [n.d.]. *Privacy Policy.* https://docs.travis-ci.com/legal/privacy-policy/

[84] Anne-Marie Tuikka, Chau Nguyen, and Kai K. Kimppa. 2017. Ethical questions related to using netnography as research method. *ORBIT Journal* 1, 2 (Oct. 2017).

[85] Norman Vinson and Janice Singer. 2001. Getting to the source of ethical issues. *Empirical Software Engineering* 6, 4 (2001), 293–297.

[86] Norman G. Vinson and Janice Singer. 2008. A Practical Guide to Ethical Research Involving Humans. In *Guide to Advanced Empirical Software Engineering*, Forrest Shull, Janice Singer, and Dag I. K. Sjøberg (Eds.). Springer London, Chapter 9, 229–256.

[87] Annie Ying. 2015. *Mining Challenge.* http://2015.msrconf.org/challenge.php