Empirical Software Engineering (2022) 27: 17
https://doi.org/10.1007/s10664-021-10057-7

Ethics in the mining of software repositories

Nicolas E. Gold[1] · Jens Krinke[1]

Accepted: 20 September 2021 / Published online: 2 November 2021

Abstract
Research in Mining Software Repositories (MSR) is research involving human subjects, as the repositories usually contain data about developers' and users' interactions with the repositories and with each other. The ethics issues raised by such research therefore need to be considered before beginning. This paper presents a discussion of ethics issues that can arise in MSR research, using the mining challenges from the years 2006 to 2021 as a case study to identify the kinds of data used. On the basis of contemporary research ethics frameworks we discuss ethics challenges that may be encountered in creating and using repositories and associated datasets. We also report some results from a small community survey of approaches to ethics in MSR research. In addition, we present four case studies illustrating typical ethics issues one encounters in projects and how ethics considerations can shape projects before they commence. Based on our experience, we present some guidelines and practices that can help in considering potential ethics issues and reducing risks.

Keywords Research ethics · Mining software repositories

1 Introduction

There have been a large number of papers that report the mining of data contained in software repositories, i.e. software data such as source control systems, defect tracking systems, code review repositories, archived communications between project personnel,

Communicated by: Georgios Gousios and Sarah Nadi

This article belongs to the Topical Collection: *Mining Software Repositories (MSR)*

✉ Nicolas E. Gold
n.gold@ucl.ac.uk

Jens Krinke
j.krinke@ucl.ac.uk

[1] UCL Computer Science, University College London, London, UK

question-and-answer sites, continuous integration servers, etc.[1] A software repository contains considerable information about the authors 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 (2008) identified that open-source software research (including data mining) involves humans as participants, collaborators, or data sources and thus requires ethics consideration.

Much research focuses on open-source software repositories and one could assume 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 permit freedom to study or do not restrict the purpose of use can support an ethics defence for studying the code without the need for further ethics considerations. 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 (The European Parliament and the Council of the European Union 2016) or the The California Consumer Privacy Act of 2018 (CCPA) (California State Legislature 2018) may be cited as sufficient. Although law and ethics are linked, they are not necessarily the same thing (Hand (2018) characterises ethics as guiding what a person *should* do, and the law as what they *must* do). Legal compliance is often considered within ethics but is not necessarily sufficient to achieve ethical safety.

Ethical practice changes in response to societal concerns, technological advances, new ethics theory, and (with particular relevance to repositories) changes in the contractual availability of data and the terms offered to those who provide it. Thus practices that were once considered ethical may no longer be so (and vice versa). As ethics theory and other factors evolve, so should ethics and research practice. For example, data ethics has only emerged as a discrete branch of ethics in recent years (Broad et al. 2017). It is therefore important to periodically look afresh at a research area to consider its current ethical situation (a point also made by Hand (2018) about data ethics in general). The Association of Internet Researchers ethics code (Franzke et al. 2020) characterises ethics as a “process” approach, identifying it as “reflective and dialogical” on the basis of ongoing experiences and reflections of researchers.

There are examples of past practices in the MSR space being revisited: email addresses were removed from the GHTorrent data dump in March 2016 (Baltes and Diehl 2016) and age information was removed from Stack Overflow’s public data as part of an audit for the GDPR (Craver 2018).

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) or legal issues (for example, Gonzalez-Barahona (2020) presented a tutorial on “Mining Software Repositories While Respecting Privacy” at the MSR conference in 2020). Indeed as a broader field, Information and Communication Technology Research as a whole is still coming to understand the breadth of ethical issues that it may need to address (Dittrich and Kenneally 2012). There are some signs that this change

[1] We interpret the term software repository in a very broad way and use the same description as the Mining Software Repositories Conference.

is taking place, e.g. Stahl et al. (2016) found several papers concerned with Internet-based research in their survey of ethics in computing. They 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 highlighted earlier by Berry (2004), suggesting resolution of the ethical matters involved is not straightforward. Research guidance like the Menlo Report (Dittrich and Kenneally 2012) 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 present ethics 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. We adopt the Menlo Report (Dittrich and Kenneally 2012) framework as the primary lens through which to examine ethics issues, and integrate relevant factors from the Association of Internet Researchers' frameworks (Markham and Buchanan 2012; Franzke et al. 2020).

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 types, we identify the ethics issues that may need to be considered in creating and/or using them (or similar datasets) in future work. We present an extended version of our initial case study on ethics issues in MSR mining challenges (Gold and Krinke 2020b) in which we focussed on the mining challenges in the years 2010–2019 (now covering 2006-2021 and incorporating the Data Showcase Track). The results of a survey on MSR research ethics views (undertaken since MSR 2020) are also presented. Having discussed issues raised by MSR data, and following indications given in the survey data, we respond by describing the ethics aspects of four case studies from our own experience, finally distilling some practical guidelines and suggested practices to support researchers in future.

At the outset, we want to establish clearly that our intention in this paper is to promote and support the development of ever more 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 in MSR research, our references to the previous MSR conference challenge and showcase 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 (given 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 the particular researchers' institutional and national requirements.

The paper's main contributions are:

1. An analysis identifying data used commonly in MSR research.
2. The results of a community survey of views on MSR research ethics.
3. Detailed discussion of potential ethics issues arising from MSR data.

4. Guidelines and practice recommendations for researchers in approaching MSR research ethics.

The paper is structured as follows. The next two sections set out the motivation for what follows, presenting a survey of the MSR mining challenges and the papers of the Data Showcase Track, followed by the results of a small survey of researchers in the community. Having thus established the need to discuss and explore ethics issues in this context, Section 4 presents the framework of the Menlo Report and how it may be applied to MSR research. Sections 5 and 6 analyse the data sets of the mining challenges and their specific ethics issues. Four case studies on how ethics considerations have shaped projects are discussed in Section 7. Section 8 discusses our observations. Some guidelines and suggested practices are presented in Section 9. Sections on threats to our research, related work, and conclusions follow.

2 Ethics Reports in MSR Mining

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 2006 to 2021 of the MSR mining challenges. We also did a simple analysis of the papers published in the years 2014 to 2021 of the MSR Data Showcase Track.

2.1 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 141 papers that were published for the Mining Challenge Track in the years 2006 to 2021. 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 4.2.3, 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 relatively recent, we found some, but not extensive, discussion: only one mining challenge (Wilkie et al. 2018) discussed ethics issues in some detail, and two papers (Soto-Valero et al. 2018; Abric et al. 2019) mentioned the anonymity of the underlying datasets in the research that built on them. 35 papers contained discussions of threats to validity (perhaps a section of a paper best suited to the discussion of ethics issues).

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

Table 1 Overview of the challenges by year

Year	Papers	VCS data	Issue tracker data	Mail Archives	Stack Overflow	Build logs	IDE events	Published papers	Threats to Validity
2006	Pinzger et al. (2006)	×	○	○				12	0
2007	Zimmermann (2007)	×	×	○				6	0
2008	Kim et al. (2008)	×	×	○				5	1
2009	Bird et al. (2009)	×	×	○				5	1
2010	Hindle (2010); Hindle et al. (2010)	×	×	×				6	0
2011	Schröter (2011a, b)	×	×					5	1
2012	Shihab (2012); Shibab et al. (2012)	×	×					6	1
2013	Bacchelli (2013)				×			12	1
2014	Gousios (2013); Baysal (2014)	×						8	2
2015	Ying (2015)				×			14	5
2016	Dyer (2013); Dyer et al. (2013, 2015); Nguyen and Dyer (2016)	×						10	4
2017	Beller et al. (2017a, b), The TestRoots Team (2020)					×		14	2
2018	Proksch (2017, 2019); Proksch et al. (2018b; a)						×	13	7
2019	Baltes et al. (2018, 2019); Diehl et al. (2019); Baltes and Diehl (2019); Baltes (2019; 2020)				×			14	7
2020	Pietri et al. (2019; 2020a, b)	×						3	2
2021	Karampatsis and Sutton (2020); Allamanis et al. (2021)	×						8	1

the corresponding year and in how many of the papers we could find a “Threats to Validity” discussion.

2.2 Data Showcase

The MSR conference also has a Data Showcase Track, introduced in 2013, to encourage researchers to share their data and to provide a forum to share and discuss data sets that underpin the work of the MSR community.

We analysed the 112 papers published in the Data Showcase Track in the years 2014 to 2021. We did not include 2013 as the Data Showcase Track papers could not be distinguished from papers in other tracks. 17 papers mention a threat to validity. One paper explicitly mentioned that the data was used with consent of the organisation behind it (Gonzalez-Barahona et al. 2015). One paper did discuss ethics issues and described how the data was pseudonymised (Fry et al. 2020). Another paper had a detailed discussion on privacy, and legal and ethics issues (Robles et al. 2014). The authors discuss the implications of combining the data from a survey and the data extracted from other sources. Moreover, they describe the limits of anonymisation of such data and how secure anonymisation could help. Rao et al. (2021) discuss how their data was anonymised and how they filter out queries that were entered by less than *k* users and could potentially contain sensitive information. Yamashita et al. (2017) discuss how the history of Git repositories was re-written to remove personal data. A dataset containing videos discussed the anonymisation of the video materials (Yamashita et al. 2018). Matalonga et al. (2019) describe how they gathered anonymous usage data about the usage of an app running on mobile devices. Six more papers mentioned that their dataset did not include user names and email addresses and/or how privacy was ensured. Markovtsev and Long (2018) discuss how their dataset complies with GitHub terms and conditions.

We will discuss some of the above mentioned ethics issues in more detail later. It is interesting to see that papers in the Data Showcase Track discuss ethics issues more often (14 out of 112 papers) than papers in the Mining Challenge Track (one out of 141) although the number of papers discussing ethics issues at all is still a relatively small proportion of the total. However, sometimes the website accompanying a paper contains discussion related to ethics, e.g. Pietri et al. (2019) do not discuss ethics considerations but the accompanying website contains an Ethical Charter (Software Heritage Archive 2019b).

2.3 Data Sources

The mining challenge started in 2006 and until 2012 the challenges were similar. In the initial four years, the challenge was to analyse open source software projects and the chosen projects changed between the years. Each year, the organiser provided copies of version control data and issue tracker data, but encouraged the participant to also use other data sources for the projects, in particular issue trackers and mail archives (shown as '○' in Table 1). From 2010 onward, the focus was on the provided data set and the encouragement was weakened to a simple "Feel free to use ..." and dropped in 2014.

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

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.

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.

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.

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.

We will discuss the six identified data sources and their potential ethics issues later in this paper.

3 Community Opinions and Experience

Having seen in our literature analysis[2] that there is often little discussion of ethics issues, we carried out a survey of those involved in mining repositories to seek information from the research community as to its perceived needs and views on ethics in MSR. The purpose was to help inform recommendations for routes to stronger ethical practice in the MSR community.

Participants were recruited in two rounds, the first using an advert shown at the end of our MSR 2020 talk (Gold and Krinke 2020a) and associated Twitter posts, the second via advertising through Facebook and Twitter channels several months later. The potential pool of recruits thus extends beyond those attending the conference. The first round response was low ($n = 5$); the second was better ($n = 12$) and therefore the combined set of 17 responses has been analysed. Whilst this response level is too low to treat the data as representative of the community as a whole, it is nonetheless interesting to consider the views and experiences conveyed.

The survey covered five areas: demographics, research ethics training and opportunities, views on the need for ethics in various MSR situations, experiences of interacting with Research Ethics Committees (REC) and/or Institutional Review Boards (IRB), and views on ways to improve ethics awareness and consideration.

Demographics Respondents' MSR experience ranged from 0–1 up to 10+ years. Eleven had MSR research as their main focus, six do MSR research but not as their main focus. All had analysed repository data prepared by others; thirteen had prepared a dataset for others to use. Regarding roles (non-exclusive categories), eleven indicated that they were researchers, and six professors/lecturers. Nine worked in academia, one also in industry and one solely in industry; five were postgraduate students (two of whom were also research assistants). Ten indicated that they had engaged with the paper (Gold and Krinke 2020b) and/or tutorial (Gonzalez-Barahona 2020) connected with mining and ethics at MSR 2020. From this data, we see a tendency in the responses towards academic research.

Research Ethics Training and Opportunities The survey asked about respondents' confidence in their knowledge of research ethics and from where they obtained this knowledge. A majority (ten) felt they had sufficient confidence in their understanding. There was no apparent correlation between experience and understanding: one of the 10+ years-experience respondents indicated a lack of confidence in ethics understanding, as did two of the 0–1 years-experience respondents. There was no obvious correlation between role and confidence in understanding either.

Nine respondents would like to increase their knowledge but do not have time, six would like to increase their knowledge but do not know how (including two who do not have time). In terms of training activity, five undergo on-demand training, nine rely on others (although six of these are happy with their own knowledge), two covered ethics in their research training, and no-one does regular training. Three gain knowledge in other ways. It

[2] The initial version of which was first reported at the MSR 2020 conference (Gold and Krinke 2020b).

is interesting to note that all participants who have obtained their research ethics knowledge as part of their research training or who have gained their knowledge in other ways than the ones mentioned in the questionnaire feel confident in their understanding of research ethics.

In terms of organisational training provision, eight respondents indicate that they do not know if their organisation provides ethics training, six organisations are reported as requiring research students to undertake it, three require staff to do so. Only one organisation is actually reported as providing training, and three are reported as not providing it. The picture in terms of organisational opportunity and communication of that opportunity is thus mixed.

Views on the Need for Ethics in Various MSR Situations Views on the ethics considerations needed for analysing pre-assembled datasets varied. Four respondents indicated that these must always be considered to check the conditions of data collection. Seven think these must be considered to check for new risks (including two of the 'data collection' respondents, and another three who consider that it depends on the analysis). Overall, ten responded that the need for consideration would depend on the analysis. Two felt that no consideration was required for different reasons: one because the data is public and anonymous, the other because creators will have given prior consideration to the issues. The former of these two respondents also indicated that the need would depend on the analysis and that new risks would need to be considered.

Figure 1 shows how respondents feel about which types of data need particular levels of ethics scrutiny. It is notable that the majority of data types are considered by the majority of respondents as potentially needing ethics consideration. Binaries are the clear exception.

Resolving ethics issues impacted projects differently. Ten respondents shaped projects as a result of resolving ethics issues, six beforehand, two during, and two in both situations. Two people shaped projects to avoid the effort of needing to apply for approval (one who had shaped for resolution beforehand), two to avoid the risk of rejection (one of these reported a project rejected at the first or subsequent attempts, and also shaped beforehand), and three on the advice of colleagues and/or reviewers. Two shaped projects for other reasons, and five indicated that they have never shaped projects for ethics reasons (although one of these had previously indicated they had done so during a project as above, thus the response has some inconsistency). Five respondents have abandoned an MSR project because of ethics issues

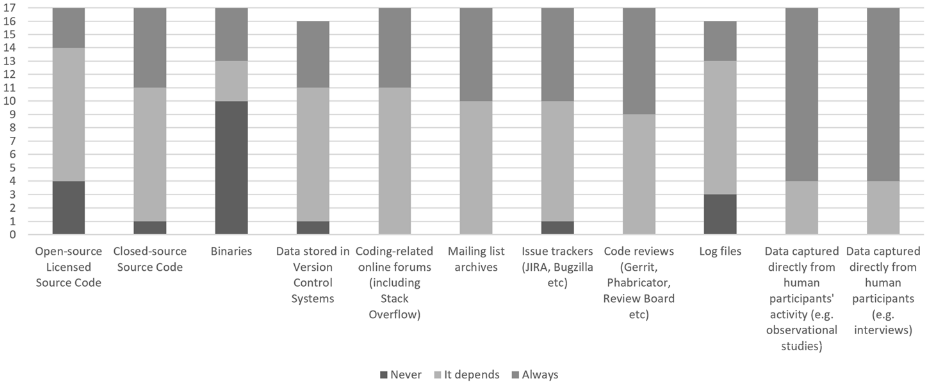

Fig. 1 Respondents' views on the level of ethics consideration needed for different types of data (note two categories contained non-responses)

(one of them more than once): one because ethical issues could not be resolved beforehand, two because of approval process complexity and the expected time needed to get approval, a further respondent because of the expected approval timescale, and one for other reasons (this respondent indicated they applied for ethics approval for reasons other than MSR). No-one had abandoned work because of rejection or because the ethics approval timescale was too long relative to the constraints on research time.

Experiences of Interacting with RECs and/or IRBs Eight respondents confirmed that their institution has a REC/IRB/equivalent (five respondents' institutions did not, and four did not know). Of the eight who have a REC/IRB to apply to, two respondents do not apply for approval (on the grounds that their research is not relevant to it), two apply rarely, and four apply regularly. Considering the six respondents in the latter two groups, five applied because of the nature of the work (the respondent who did not indicated that they rarely apply), and in one case because of national requirements. Only one person had an organisational requirement to submit all work for review. One regular applicant does so because their organisation requires this kind of work to be submitted, but they disagree with the position. One applied for other reasons. None indicated publication venue requirements for ethical approval, and none applied because a colleague thought they should.

The majority of those interacting with RECs/IRBs felt that there were issues of understanding the research: No respondents felt that their REC/IRB understood the problems around repository data, three felt that their REC/IRB did not understand the properties of repository data, and one felt that their REC was unhelpful. More positively, two respondents felt supported by their REC/IRB and one received helpful comments. One respondent indicated that their interactions related to work not involving MSR.

The impact of ethics approval on research design, outcomes, and impact was marginal at best. In terms of design, three indicated no difference, two positive, and one negative. In terms of outcomes, one positive indication was returned, and five indicated no difference. In terms of impact, four people indicated no difference, one negative, and one positive (interestingly, the same person who had indicated that approval had negative impact on their research design).

Views on Ways to Improve Ethics Awareness and Consideration The final questions asked for respondents' views on a range of ways in which ethics practice in the MSR community might be encouraged and/or developed (these were suggestions we listed in the survey). Table 2 shows them in descending order of support. The highest support was seen for suggestions focussed on incorporating ethics in peer review, and for the creation of guidelines and training (shown above the stronger dashed line). These were the only two suggestions that received more than 50% support among the respondents. The next four (down to the weaker dashed line) most-supported suggestions are also in the area of guidance, training, and process. Those adding requirements to existing publication processes received less support.

It is interesting that the most support falls to those suggestions relating to guidance for (and by) the community itself and to help ethical oversight bodies to better understand the nature and risks of MSR research. Since these suggestions relate to the principles of ethical research practice in this area, they are likely also to have the highest impact on the widest range of related activity.

Some of the support differs between the two groups of participants who feel confident in their understanding of research ethics (10) and those who do not (7). The participants

Table 2 Support for ethics practices in the MSR community (practices receiving more than 50% support are shown in bold)

Practice	Support	Confidence	
	(of 17)	Yes	No
Peer reviewers should consider ethics issues	**11**	**8**	3
Create ethics guidelines by and for the MSR community	**10**	**7**	3
Encourage prof. bodies to publish ethics guidance specific to MSR research	8	4	**4**
Publication venues should require an ethics statement in all papers	**8**	6	2
Create ethics guidelines to help REC/IRB bodies better assess MSR ethics risks	7	3	**4**
Run regular ethics sessions at annual conferences	7	3	**4**
Suggest adding licences to cover repository metadata as well as code	6	3	3
Require a discussion of the ethics considered in every paper	6	4	2
Publication venues should require a statement of REC/IRB approval	5	3	2
Publication venues should require proof of REC/IRB approval	2	1	1
Maintain repositories for research project announcements to permit withdrawal	1	1	0
Other	1	1	0
Nothing	0	0	0

who feel confident strongly support the idea that peer reviewers should consider research ethics (8/10) and the creation of ethics guidelines by and for the MSR community (7/10). Moreover, the majority of them (6/10) also support the suggestion that publication venues should require an ethics statement in all papers.

Discussion The low number of respondents makes it impossible to draw substantive conclusions and the data is likely skewed as the participants may have a heightened interest in ethics, having mostly attended our presentation and participated in the survey. The most that can be said is that there is some variability in awareness of ethics issues among the respondents but a willingness to engage with them where topics of research demand it. In some cases more training is desired, and where possible, more support from the community both for researchers and for REC/IRB bodies to increase understanding of the particular issues involved. This is reflected in the relative support for ways in which to improve practice, with the most well-supported suggestions relating to ethics guidance and peer review. This paper attempts to give some ethics guidance by the discussion in the following sections including the set of guidelines and good practices in Section 9.

4 Ethics Frameworks for MSR Research

Our literature analysis showed that there has historically been relatively little discussion of ethics issues in published papers at MSR conferences. Our subsequent survey offered a degree of anecdotal support for the creation of ethics guidelines for the community and for ethics oversight bodies like RECs and IRBs. The need for guidance was also identified along with the need for an accepted code of ethics for analysing software data in a panel at the International Conference on Software Engineering in 2014 (Menzies et al. 2014). Following our analysis here, we offer some suggested guidance in Section 9 later.

4.1 Balancing Ethics Issues in Research

A fundamental question for researchers is how they can defend the ethical nature of their research to those who have oversight and to the public in general. This requires articulating the benefits, risks, and controls. Whitney (2016) argues that in recent years the necessary balance between the welfare of research subjects and the general desire to create scientific or societal benefits has been lost, with IRBs leaning strongly in favour of subject protection above all else. Whilst subject protection is important, Whitney makes a strong case for balance between the two aspects. Hand (2018) gives a privacy-related example of the need to consider balance: protecting the privacy of someone suffering a fatal and highly contagious disease vs. the protective societal benefit of revealing this. Achieving such balance can be significantly aided through the use of a framework of ethical principles appropriate to the disciplinary area concerned. The wide range of potential investigations based on repository data means that applicable frameworks may vary.

The way in which discussions of balance are framed may be affected by the intellectual and philosophical traditions that underpin research. For example, the Association of Internet Researchers (AoIR) ethics code (Franzke et al. 2020) describes European and Scandinavian approaches as strongly deontological, and UK and US approaches leaning more towards utilitarian (teleological) approaches. We do not adopt either of these positions in what follows, preferring instead to follow the AoIR principle of ethical pluralism. Thus although we discuss factors in balancing ethical risks with research outcomes, the determination of that balance in specific circumstances is a matter for individual researchers and their oversight bodies in the contexts in which they work.

Ethical theory has a long and rich history: Stahl et al. (2016) clearly and concisely summarise the various philosophical and ethical theories that underpin the spectrum of ethics assessment in computing in general, identifying the ethics of computing as a component of applied ethics. Codes of IT professional ethics, e.g. the BCS (BCS: The Chartered Institute for IT 2021) or the ACM (Association for Computing Machinery 2018), typically do not address research matters in detail (although the ACM code does so to some extent in its illustrative examples), focusing more on general professional conduct. Of course, it can be argued that professional conduct for researchers encompasses the management of research ethics but a more explicit treatment of ethical principles in research is helpful.

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 (The British Psychological Society 2018, 2014), BERA2018 (British Educational Research Association 2018), RESPECT (Dench et al. 2004), the Menlo Report (Dittrich and Kenneally 2012), and the Association of Internet Researchers guidelines (Markham and Buchanan 2012; Franzke et al. 2020). 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).

The following sections adopt the Menlo Report (Dittrich and Kenneally 2012; Bailey et al. 2012) as the primary ethical frame since this is a broad framework for research in the Information and Communication Technology space and thus can expose the set of issues to be considered in a relevant way. We augment the discussion with considerations from the two most recent AoIR ethics codes (Markham and Buchanan 2012; Franzke et al. 2020) that account more explicitly for the internet context of much MSR research and some of the issues that arise as a result.

4.2 The Menlo Report

The Menlo Report (Dittrich and Kenneally 2012) was published in 2012 with the intention of providing a set of principles to help researchers identify and manage ethical issues in what it defines as Information and Communication Technology Research (ICTR): *"...involves the collection, use, and disclosure of information and/or interaction with this ubiquitously connected network context which is overlaid with varied, often discordant legal regimes and social norms"*. Its secondary aim was to support those involved in assessing and authorising research (IRBs, RECs et al.). The framework was further explained through application to various case study scenarios (Dittrich and Kenneally 2013). Although the case studies do not encompass MSR research specifically, the framework strikes a good balance between general ethical principles and their application to ICTR.

The Menlo report identifies various challenges associated with undertaking ethical research, for example, indirect interaction with humans arising from increased logical or physical distance between the researcher and those affected by the research, the ease with which large numbers of human subjects (or their data) can be accessed through systems, and the associated increase in speed and scope of impact for potential harm. It also explicitly acknowledges the complexity of the legal and social environment that frequently spans multiple jurisdictions and countries. The main reasons that the Menlo Report is well-suited to ethics analysis in MSR research include its explicit acknowledgement and requirement to consider a very broad range of stakeholders (including indirect "participants" and platform owners) and the computer systems that support those individuals who are not research subjects themselves.

The Menlo Report sets out four principles for the consideration of ethics, three from the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979) and the fourth added for the evolving nature of the legal contexts of privacy, and information and systems assurance. A summary of the principles and how they are applied is shown below:

1. *Respect for Persons* (voluntary participation following informed consent; respecting individuals as autonomous moral agents 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* (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 Menlo 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.

4.2.1 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. Platform owners may also be impacted, particularly if large-scale mining is undertaken that could impact their service. 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 which, aside from practical matters around access impacts, may wish to be consulted before agreeing to the use of its users' data for research (since it is the gatekeeper to its communities), 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 researchers to follow responsible disclosure principles), or MSR research on developer behaviour could accidentally uncover malicious behaviour by a developer.

4.2.2 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?

Ethical consent may be sought directly, or may be regarded as implied through terms and conditions or licence application as discussed further in Section 5. Direct ethical consent can carry with it the conditions that the researcher chooses to offer (e.g. withdrawal of data up to a certain time after collection, no withdrawal of data and so forth; the participant can then decide if they are happy with the offered conditions before choosing to participate). Indirect/implied ethical consent may need to consider withdrawal differently, perhaps on the basis of continued public visibility or otherwise of the data being analysed.

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 (Benbunan-Fich 2017; Sugiura et al. 2017) (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. To appropriate contributors' activity data in this way could be considered an affront to their autonomy. This is different to the code-study situation where a developer's application of a licence offsets some of the ethics concerns.

Terms and conditions are not always a panacea in this case either as various problems arise: difficulty in determining who to approach for consent (Berry 2004), the need to actively consent rather than passively agree terms without reading them (Hand 2018), 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 (Benbunan-Fich 2017), and the changing participation in open-source developer communities meaning that it may not be possible to seek consent from individuals (Oezbek 2008).

Issues of consent in internet-based research are the subject of considerable ongoing debate and there are a wide range of positions (that themselves build on different intellectual and philosophical traditions) around the extent of public access vs. expectations of privacy, the control that platforms have over their environments and interactions vs. the need to research them, and the enforceability of terms and conditions.

Where the Menlo Report identifies the need to consider these issues, the AoIR ethics frameworks (Markham and Buchanan 2012; Franzke et al. 2020) and accompanying companion essays, e.g. by Locatelli (2020), address them in more depth and from a pluralistic viewpoint. The AoIR framework (Chapter 3: Internet Research Ethics 3.0; Franzke et al. 2020) itself identifies the difficulties (and resulting serious ethical dilemma) of obtaining informed consent in big data projects, particularly where automated data collection is used. Approaches to handle this that are reported in the AoIR framework document include: seeking informed consent directly, anonymisation (extremely hard for MSR research), careful justification of sensitive data processing, and deferring the seeking of informed consent to the dissemination stage (i.e. seeking direct consent from those whose data may be involved in a publication; this may be hard for MSR given the scale of processing often involved). Note that the thrust of the AoIR framework is pluralistic and reflective so whilst these approaches may be good to consider, they are not necessarily solutions that can be applied without detailed thought and justification.

4.2.3 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. 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 potentially 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 (Software Heritage Archive 2019b) 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 (Software Heritage Archive 2019a) and load management on its own server, e.g. disallowing massive-scale data extraction in order to equitably serve its users (Software Heritage Archive 2019c). 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 research (and software research generally) may relate to observations and judgements of practices, impacting on the researched 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 (Oezbek 2008). 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.

4.2.4 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 may be somewhat harder to attain however. In research involving people, fair selection is often addressed using random sampling within a population. However, it is important that research is designed to be inclusive in terms of both questions and process, and that potential participants are not excluded for reasons other than the specific study exclusion criteria that apply to all. 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. Moreover, research often focuses on repositories from a limited number of communities and using random sampling is therefore restricted to the members of such communities. Bosu and Sultana (2019) investigated the gender diversity in open source software and found that the lack of gender diversity remains an ongoing issue and the ratio of female contributors to open source software is significantly smaller than the ratio of females in computer science.

4.2.5 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 (Dittrich and Kenneally 2012) and the fact that legal compliance is part of ethical data handling (Broad et al. 2017). 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) (The European Parliament and the Council of the European Union 2016) 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 (2020). 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 (directly) what might be considered 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 (The Linux Foundation 2018) 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). There are other legal bases that can be used in different circumstances, including consent. Consent as a legal basis for data processing under the GDPR is different to ethical consent and carries with it particular legal data subject rights. Of particular note for researchers using 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 (The Linux Foundation 2018).

Withdrawal of GDPR consent is the topic of ongoing discussions on mailing lists and forums in the area of repositories, in particular in version control systems. Withdrawal of consent (under GDPR) should lead to the deletion of a person's data (where mitigating conditions do not apply). This can be achieved by the deletion of the person's contributions or by deletion of the personal data like name, email address etc. While in some repositories this is straightforward, in Git repositories this cannot achieved without rewriting the complete history of a project (Yamashita et al. 2017). Moreover, some argue that deletion of such data should not be allowed in order to demonstrate the provenance of contributions and intellectual property. Another challenge is that withdrawal and deletion from the original data may require the withdrawal and deletion from datasets derived from the original data. This could have potential implications for publications based on the data since the publications may no longer be able to refer to the data if consent was withdrawn in the future (although as noted, there are some limits to the right to erasure under certain circumstances that may mitigate this; legal advice will likely be required in these situations).

Although significant in impact, the GDPR and the CCPA (and other data protection legislation) are not the only legal frameworks that may apply (e.g. Broad et al. (2017) also note laws around confidential information and anti-discrimination as applicable to data handling). In the US, the HIPAA legislation (U.S. Department of Health and Human Services 1996) 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. (2017) note, database law). Copyright law is the foundation of open-source code licensing. Since the intellectual property rights in source code are usually held by the author, licences are required to make it legitimately obtainable by others (and corresponding contribution licences required to assign rights to the distributing projects). There may be legislative exceptions for the use of unlicenced code (e.g. provisions for fair use or fair dealing, but these may be restricted by case law and may limit the extent of the code that can be used). Other aspects may support purposes beyond the licence terms (e.g. the text and data mining exceptions in UK copyright law, although these require legitimate access to the material being mined and that may in itself still require a licence).

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” (Free Software Foundation 2021) or “open source” (Open Source Initiative 2019) software. Compliance with these conditions generally permits the unrestricted use of licenced work for any purpose (see the lists and associated conditions for more information). 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 (Open Source Guides 2019). 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. Some licences (e.g. Apache License 2.0) give more coverage, encompassing all contributions to a project under the licence. Interestingly, at the time of writing GitHub’s general terms and conditions would appear to cover all contributions of any kind to a repository under the default licence (which applies where another licence has not been adopted, the default licence being generally more restrictive in terms of content use than a typical open source licence), but once a licence has been adopted, that is the licence that applies. This may thus differ for the same repository hosted on another site.

The scope of compliance extends beyond laws and licensing to the terms and conditions of services and platforms that are used for research purposes. For example, outside of the MSR context but of relevance and within the scope of investigations covered by the Menlo Report framework, it was recently reported that Facebook has asked a group at New York University to cease collecting data from its site on the grounds that it has not given permission for the collection (Horwitz 2020), leading to an interesting discussion of the ethical balance between research needs, data subject rights, and platform rights (Naughton 2020).

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 (Thomas et al. 2017) 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 (2017) 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.

5 Ethics in Mining VCS

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. Only three out of the 15 mining challenges did not use such data. Given the relative occurrence of such data in studies, we discuss it at greater length than the other types covered in the next sections.

5.1 Studying Code

We first discuss the ethics of mining the code contained in version control systems. Such data is often hosted on and obtained from public repository sites like GitHub, GitLab or BitBucket. Those sites may impose terms of access to the repository data obtained through them and these may themselves interact with the licences applied to the code. As such, there are a set of complex ethics issues around user expectations and agreements, and the available evidence that might be used to defend a proposed piece of research to a REC or IRB.

As noted in the introduction, the code in a repository typically has a licence applied to it by the author(s). The terms of such licences vary but those compliant with the FSF and/or OSI licensing criteria permit study (or do not restrict use) and therefore permit research without further reference to the authors. Since the hosting sites make this available and the data is licensed for use, in respect of source code itself (and any data contained therein) there would appear to be few, if any, ethics issues involved in studying that code (however, using the code as a way to study the developers would be a different situation and would likely raise both data protection and ethics issues).

One possible complication might arise in repositories where code has changed licence at some point in the development history. In that situation, the act of cloning the repository (with the full history included) would result in a researcher obtaining data under multiple licences (including none), some of which might be incompatible with the purposes of research and thereby re-raising concerns about developer expectation and consent.

It may be helpful to consider whether or not a particular piece of code has been ‘manifestly published’, i.e. it has had a licence applied that permits publication and study. If so, that would place it in a similar situation to a published manuscript (where the process of manifestly publishing involves sending the work to a publisher). However, code that has no licence applied may be viewable by the “public” on a hosting website but it cannot be safely assumed that it is legitimately available for use if there is no evidence of an act of manifest publication (a developer may simply have made a mistake in their repository privacy settings). For example, at the time of writing GitHub allows projects to adopt a licence, which is then visible on the project’s GitHub page and can be retrieved via the GitHub API. As long as a project has not adopted a licence, the rights granted to users are limited to “*use, display, and perform [the content] through the GitHub Service and to reproduce [the content]* *solely on GitHub as permitted through GitHub's functionality (for example, through forking).*" (GitHub Inc 2020a). The terms also establish a default equivalence between inbound and outbound contributions. By contrast, the terms of Atlassian Bitbucket *require* licences to be applied to any public repository and these must cover inbound and outbound contributions appropriately. Whether licences are adopted (and if not, what is implied from the terms), which licences are adopted, and for what purpose, thus become important points of information when developing consent arguments based on them. In general one cannot use, copy, etc. work (here repository content) that has no licence (unless exceptions apply, see Section 4.2.5). Therefore the usage of projects without a declared licence in MSR research (or research in general) may need particular scrutiny. In the absence of any licence, a researcher will need to consider the developer's expectations and any other evidence of consent to study. In the case of GitHub one has to consider that a repository owner may have created a public repository only because private repositories required payments until 2019. Therefore deriving consent from the argument alone that a repository is public may be seen as problematic.

Limiting research to projects with a declared licence compliant with the FSF and/or OSI licensing criteria is perhaps the simplest approach (since these permit study and/or do not restrict the purpose of use). Other licences would need to be considered on an individual basis for what they permit. However, restricting MSR research to FSF/OSI compliant repositories may introduce potential bias in the same way that MSR research usually applies selection criteria to create subsets that are analysed and where the results can only be generalised to projects that fit the selection criteria. Research that needs to generalise, e.g. to study licensing, has a reason to not restrict the research but would need to consider the ethics implications of doing so.

5.2 Using Commit Records

The code situation is complicated by the data contained within the commit records of the repository. Cloning a repository provides this data to a researcher (and may require data protection steps to be taken since it is personally identifying) even when their research is not related to it. Even if the code is licensed, such licences do not typically apply to any other type of data. It could therefore be argued that this is metadata and requires separate consent considerations for use (posing challenges to those who cannot avoid acquiring it in cloning, but actually do not need it for their research). One might argue that since developers can obscure their identities in commits, they can avoid being studied if they so wish, however, this seems ethically suspect since that would require a developer to modify their behaviour and ability to receive credit for their work in the context of their development community, simply to avoid the possibility of a researcher wanting to use their code.

A more ethically defensible position derives from the principle of manifest publication. Whilst one cannot easily argue that simply providing commit metadata is itself evidence of intent to publish that data, one can draw a parallel with the process of publishing a paper. In that situation, author name, affiliation, contact information, and frequently an abstract are provided separately to the manuscript itself. These are published alongside the manuscript to enable searching and indexing. Commit metadata can be seen in the same category as this data (when used in relation to retrieving the associated code): it is data arising from the act of manifest publication and accompanies the publication itself for the purposes of attribution and retrieval. One might therefore seek to justify the use of the equivalent commit data for code-related research on similar grounds: it is data pertaining to the process of publication and supplied for that purpose and in the full expectation that it will accompany the published work when that work is retrieved for use.

In a similar way to using code to study developers, using commit metadata to study developers themselves would raise ethics issues that would need to be addressed: that is not the same use as having that data in order to enable the study of code since it would be being used as a proxy for developer activity, not a means of access to non-sensitive material.

GitHub has recently consulted on a proposal to clarify its Terms of Service such that where a project has a notice of licence, that licence extends to the content of the repository also (including issues and comments) (GitHub Inc 2020b). It is unclear how effective this would be since some licences are particular about what they licence (e.g. code and documentation) and the terms in a licence may not be sufficiently well-defined (e.g. is documentation user-facing or developer-facing or both?) to make this application straightforward.

In contrast to the discussion above, some licences explicitly extend to project contributions beyond source code. The Apache License 2.0 (The Apache Software Foundation 2004) explicitly includes all contributions to the project, where contributions are defined as "*any form of electronic, verbal, or written communication sent to the Licensor or its representatives, including but not limited to communication on electronic mailing lists, source code control systems, and issue tracking systems*." Therefore, projects adopting the Apache License 2.0 automatically include repository data in the licensed content.

In spite of the potential difficulties, such a change to licensing policy might ease the compliance aspects of research ethics but may still pose some challenges to participant protection and consent to research. Since software licences are based on copyright, there is also a question of whether such a licence is appropriate for all content in a repository. It may not all be considered copyrightable, and if it is, then the person putting that content into the repository would need to know that they have the right to what they are supplying there, and the right for it to be sub-licensed via the software licence. Quoting material into a discussion forum from a copyright source might be considered as fair use/dealing if that is the only place in which it exists, but sub-licensing it to every copy of the repository everywhere may go beyond what a fair use/dealing justification can support and thus leave the original poster of the material in difficulty. As we note below, clarifying the licensing of repository data (as distinct from code) is certainly a good direction to go in, but it may be that such licences need to be specific and bespoke to the data themselves.

5.3 Menlo Principles Applied to VCS Data

After the general discussion above, we apply the principles of the Menlo Report to consider ethics when using version control system data.

Identification of Stakeholders and Informed Consent Stakeholders in version control systems will be (at least) the developers who contribute code, anyone contributing to issues lists or bug repositories, the platforms hosting the repositories, and those who use those repositories and software in their activities.

As discussed in Sections 5.1 and 5.2, obtaining direct prior consent from developers contributing to version control systems is usually difficult to impossible (although projects or platforms that require explicit agreement to a Contributor Agreement make a form of "pre-consent" easier to demonstrate based on having the terms of inbound contribution clearly defined, rather than needing to rely solely on the implications of the outbound licence terms of the software). Another 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 (Creative Commons 2009) licence. Whilst the data may be licensed, there may still be specific risks raised by particular research that require 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 ((Dittrich and Kenneally 2012, pp10–11) discuss this in the Menlo Report). 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 (Yamashita et al. 2017). 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 (Baltes and Diehl 2016).

Fairness and Equity Intentionally inequitable selection of subjects is unlikely as the metadata contained in repositories does not usually include sensitive information like gender or religion. The communities hosting repositories, however, often have such information. For example, the ongoing OpenStack gender diversity research uses such information (Cortázar et al. 2018). Other gender diversity research projects have used tools and other data resources to identify the gender of contributors, for example Bosu and Sultana (2019) investigated the gender diversity in open source software by mining code review repositories and Vasilescu et al. (2015) used GitHub projects to study how gender and tenure diversity relate to team productivity and turnover. Working with sensitive research topics requires considerable ethical care, particularly where tools are concerned, since automated judgements may carry unintentional bias, or may incorrectly ascribe characteristics to individuals (a significant risk to individuals if they are identifiable, and broader risk if statistical judgements are made about the state of a field and actions taken as a result).

Compliance, Transparency and Accountability Given the foregoing discussion of consent and the heavy use of compliance with the licence and access terms as a route to evidencing consent, there is little more to add here in respect of VCS data. Consent arguments based on terms and conditions will inevitably involve a researcher in detailed consideration of compliance issues as part of developing their ethics position.

However, it is important to check and comply to the terms and conditions of the organisations providing the repositories. For example, using email addresses gathered from GitHub for sending out invitations to participate in research would not only raise ethics concerns, but would also appear to violate GitHub's terms and conditions:

> *"You may not use information from the Service (whether scraped, collected through our API, or obtained otherwise) for spamming purposes, including for the purposes of sending unsolicited emails to users..."* (GitHub Inc 2021)

Indeed, researchers have been told by GitHub's support team that using email addresses in this way is not permitted[3].

6 Mining Non-VCS Data

We now discuss the different types of datasets listed above which are not derived from version control systems and the ethics issues such data may raise in future.

6.1 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). The associated PhD thesis (Proksch 2017) 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).

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 to 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 pressured 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.

[3] https://github.community/t/use-public-email-address-to-send-research-survey/ (last accessed 1st July 2021)

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 (Proksch 2019) states that "the captured feedback structure was discussed with the privacy council of a large German IT company and complies to German privacy requirements."

6.2 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 (Travis 2019).

Balancing Risks and Benefits The risks in this type of data largely revolve 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.

6.3 Mining Stack Overflow

Stack Overflow is the go-to Q&A website for programmers. Stack Exchange (the organisation behind Stack Overflow) 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 licence 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 (Atwood 2009), 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 (Craver 2018). Lin and Serebrenik (2016) report on applying different "gender guessing" approaches to Stack Overflow data.

Compliance, Transparency and Accountability The CC-BY-SA Creative Commons licence 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 of violating licences as Stack Overflow is known to contain code fragments that potentially violate their original licence (An et al. 2017; Ragkhitwetsagul et al. 2019).

6.4 Mining Issue Trackers

Data from issue trackers come in various forms, for example, they are aggregated in the Ultimate Debian Database (Debian 2020b) 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 in the terms and conditions or privacy policies. For example, Debian's Privacy Policy (Debian 2020a) 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.

6.5 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 (The FreeBSD Project 2012) 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.

6.6 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 licences 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).

7 Case Studies

The discussion of potential ethics issues was limited to six different data sources that were used in the mining challenges of the years 2006–2021. The discussion was therefore limited to a subset of data sources which are used for mining of software repositories research. In the following we expand the discussion with four case studies from the authors' own experience. They highlight the considerations of potential ethics issues before a project commences and how the considerations shape projects.

7.1 Context

The case studies discussed in the following need to be seen in the context and time in which they were done. They were all located at University College London (UCL) in the Department of Computer Science. UCL has specific regulations and policies regarding research ethics and a central Research Ethics Committee for all UCL staff and students. The overarching policy at the time of writing is:

> "All research that involves living human participants and the collection and/or study of data derived from living human participants undertaken by UCL students or staff requires ethical approval to ensure that the research conforms with general ethical principles and standards." (UCL Research Ethics Committee 2020)

There are currently a few exemption criteria that define when such research does not need central REC review and Heads of Department have final judgement as to whether a particular activity should be exempt from the requirement for central REC review in accordance with these (this does not negate the need for good ethical practice or local review: it is a risk-based triage to appropriately balance risk with review effort, higher risk proposals require more review). Obtaining approval from Research Ethics Committees in general takes time (Vinson and Singer 2008) and if projects can be shaped so as to fall under the exemption criteria so that only Head of Department's approval is necessary this can lead to faster approval (and perhaps more importantly, thereby find ways to reduce the ethical risks of the work while retaining the benefits). In UCL Computer Science, staff and students apply for exemption by consulting with a dedicated departmental Research Ethics Committee that advises applicants and the Head of Department on ethics issues and the appropriate routing for applications. In addition to first-stage review and advice, the committee provides training materials to staff and students. All students are asked to undertake this online ethics training prior to project work (the training has a particular focus on computer science project work and the kinds of issues this raises).

The first two case studies were led by Krinke without Gold's direct involvement, the third led by Krinke with Gold involved as the relevant module's second examiner and providing informal ethics advice, and the fourth led by Gold in close collaboration with Krinke as part of the research reported here.

Processes and criteria vary across nations and organisations and therefore the considerations presented in the following case studies need to be seen in the context of the discussion above and the policy frameworks under which they took place.

7.2 Case 1: Creation of a Code Review Dataset

The first case study project's aim was to create a dataset of code reviews and the state of the software projects at the time of each code review to allow investigations about how code review impacts software projects. Not only was the dataset directly used in multiple projects (Paixão et al. 2017, 2019; Han et al. 2020; Paixão et al. 2020), but it was created as a curated and reusable dataset that could be made publicly available (Paixão et al. 2018). The case study serves as an example of ethics considerations when creating a dataset for public reuse.

CROP, the Code Review Open Platform, is a curated code review repository that links review data with isolated complete versions (snapshots) of the source code at the time of review (Paixão et al. 2018). CROP currently provides data for 8 software systems, 48,975 reviews and 112,617 patches, including versions of the systems that are inaccessible in the systems' original repositories. The creation of the dataset involved mining the Gerrit repositories for the Eclipse and Couchbase communities, downloading the snapshots of the systems for each mined code review, and then linking them to create the dataset.

Mining code review platforms like Gerrit repositories raises potential ethics issues. Most of the potential ethics issues are similar to the ones already discussed. However, code review is usually tightly integrated with version control systems to allow reviewed changes to be easily merged with the code managed by the version control system. Moreover, projects often require all changes to be reviewed before being merged.

In a system like Gerrit, the code reviews are linked to commit ids and commits that have been reviewed are linked to their code reviews. Therefore, the commit metadata of projects using Gerrit not only contain information about the commit as discussed above, but also metadata of the code review linked to the commit which may include names and email addresses of the involved reviewers. Moreover, the change id allows the retrieval of the discussion of the reviewers including their personal data (if the corresponding Gerrit repository is accessible). At the time of writing, GitHub lists more than 30 million commits that contain a direct link to a Gerrit code review.

When creating the CROP dataset, potential ethics and legal issues have been considered. We will discuss them within the Menlo Report framework, similar to the discussions above.

The dataset was extracted from data that anyone can access without needing to gain permission from someone, the terms and conditions of the two communities did not restrict the use of the data, and no further analysis or profiling of the extracted data was done. In this case the creation of the dataset did not require a REC application (in the context of our local regulations and advice at the time). Details of the considerations are discussed below.

Identification of Stakeholders and Informed Consent As discussed already, it is usually difficult to impossible to obtain consent for version control systems and the same applies to code review repositories. In the case of the CROP dataset two communities are governing the repositories. We have discussed the Eclipse Foundation terms and conditions and their relevance to version control systems above, and the same considerations extend to the Gerrit repositories. The situation for Couchbase is a little bit different as Couchbase is a commercial entity that is governing the open source projects of the Couchbase community. Anybody who contributes to the development needs to sign a Contributor Agreement first so that the Couchbase open source projects can be licensed under the Apache License 2.0. As discussed above, the Apache License 2.0 is explicit in its coverage and application beyond source code and it is safe to assume that the licensing extends to the Couchbase code review data. Signing the Contributor Agreement can be interpreted as consent that the use of the code review data is not restricted in the same way that the use of the source code is not restricted.

The website providing the CROP dataset has a section on data protection which allows users to opt-out and their data be removed from the dataset.

Balancing Risks and Benefits As reviewers express opinions about the changes they review and assess their quality, the risks to them and the change authors are somewhat higher than in the case of version control systems. However, as reviewers and authors are aware of their reviews being open and public, they will be aware of the risks and use the reviewing system in a way that limits their risk.

However, the CROP dataset uses pseudonymisation to limit the risks coming directly from the dataset. All names in the collected metadata and the code review discussions were replaced by randomly generated pseudonyms. All email addresses in the code review discussions were made anonymous. As long as the actual code review is still publicly available, it would be possible to reconstruct the original name and email address. However, Gerrit repositories allow the deletion of reviews and projects often used the ability to delete reviews to clean the content of the code review repositories.

The CROP dataset also contains snapshots of the systems' source code at the time of the code review as stored in the underlying git repository. However, the snapshots do no longer contain the original commit metadata. The CROP metadata still contains the commit id of the original commit.

Fairness and Equity The metadata contained in the Eclipse and Couchbase repositories does not usually include sensitive information like gender or religion.

Compliance, Transparency and Accountability The terms and conditions of the Eclipse and Couchbase communities were analysed at the time of the creation of the dataset in 2018 (the terms and conditions may have changed since then). In particular the compliance of the CROP dataset with the GDPR was important and we sought and followed institutional advice.

The snapshots contained in the dataset are released under their original licences and the original licence headers and files are retained.

However, a first step in the considerations was to check the terms and conditions of accessing the data as provided by the two communities and the licensing terms of the repositories. The licensing (and the terms and conditions) for the Couchbase repositories under the Apache License 2.0 does not restrict the use of the repository contents in way that would have affected the creation of the CROP dataset. As already mentioned, Eclipse repository content is subject to the Eclipse Foundation Software User Agreement (Eclipse Foundation 2017a), but the repositories themselves are subject to the Eclipse.org Terms of Use (Eclipse Foundation 2019) and the content of the repositories is licensed under the Eclipse Public License (Eclipse Foundation 2017b). Again, the creation of the CROP dataset was in line with the Eclipse Foundation terms and licences.

Summary The primary risks identified and addressed in this situation were:

- Consent: implied consent justified on the basis of evidence in licence terms for contributed content.
- Privacy: for raw data, contributors are aware of the public nature of their contributions. Minimising the risk from the dataset itself is managed through pseudonymisation to make it harder (although not impossible) to recover identities.
- Focus: the publication of a dataset brings more specific attention to the data contained within it. The controls for privacy and reputation mitigate this risk.
- Reputation: since reviews necessarily speak to the quality of work done, there is a reputational risk to change authors. This is mitigated by the explicitly public nature of the raw data.
- Compliance: since the data is released under terms of licence and specific terms and conditions, these were checked to ensure compatibility with the intended study, thus also addressing intellectual property rights. Data protection procedures in force at the time were followed.
- Misconduct: failure to consider the ethics could leave researchers open to allegations of misconduct.

7.3 Case 2: Reviewer Recommendation

The aim of the research underlying the second case study was to investigate whether the inclusion of dependence data retrieved from reviewed source code could improve automated code reviewer recommendation for changes that need code review. It was an undergraduate research project in the 2019/20 academic year. The case study serves as an example of ethics considerations when creating and analysing a dataset.

To contrast ethics issues in data collection with ethics issues in data analysis, the second case study is about a project in which Gerrit repositories together with their linked Git repositories were analysed in order to recommend a suitable code reviewer for a change based on data about reviewers and authors from past changes. This is a problem that has been the focus of many previous papers (Badampudi et al. 2019).

The main difference to the first case study about creating a code review dataset is that this project not only required to collect data which was analysed, but that it also needed to identify and profile individual contributors. For example, the dataset created in the previous case study could not be used as it no longer allows to distinguish individual contributors in the source code. Therefore, this case study had to repeat the ethics considerations about code review data.

The main ethics issue is that reviewer recommendation needs to distinguish individual contributors (authors and reviewers) and map them to their individual contributions in previous commits and previous code reviews. Moreover, it requires to profile individual contributors in terms of suitability for reviewing changes of certain features. As the code review recommendation itself can neither be anonymous nor pseudonymous, it must be considered whether the datasets that are used during the project can be anonymised or pseudonymised. Anonymisation is clearly not possible as original contributors can be reconstructed. In theory, pseudonymisation would be possible by completely rewriting copies of the version control system and the code review repositories (Yamashita et al. 2017). However, such rewriting is impractical and further research is needed into how pseudonymisation of linked datasets can be achieved (Gonzalez-Barahona 2020).

In the end, the ethics considerations led to the requirement that the project needed to go through the process of applying for approval by the university's REC. The application needed to be revised in response to the received feedback from the REC to clarify to what extent contributors to the three organisations consent to the storage and analysis of their personal data. This was helped by the fact that all contributors have accepted Contributor Agreements which clarify that all their contributions are made publicly available (and open source) and that the applied software licences do not restrict the use of the data that would prevent research on the data. It was also necessary to point out that the personal data that is gathered from the contributors when registering with the organisations is held separately from their contributions and is not publicly accessible.

Identification of Stakeholders and Informed Consent The repositories that were used in the project are the same as in the previous CROP dataset case study. Therefore, the same considerations applied here. In addition, code review studies have often used projects from the OpenStack community and we planned to use code review data from the OpenStack projects, too. Developers wishing to contribute to the OpenStack projects need to become members of the Open Infrastructure Foundation. One of the guiding principles of the Open Infrastructure Foundation is *Open Development* (Open Infrastructure Foundation 2018) which makes it explicit that "*everyone can see everything about development activities, without even needing to sign up to a service*." Moreover, the OpenStack projects are licensed under the Apache License 2.0 which extends to all contributions including code review data. Similar to the discussion about Couchbase open source projects, becoming a member of the Open Infrastructure Foundation can be interpreted as consent that the use of the code review data is not restricted in the same way that the use of the source code is not restricted.

Although there is no direct consent from the participants (contributors) to use their data, all contributors have accepted Contribution Agreements which make clear that all contributions are licensed to the respective organisations and that the data may be made publicly available. They therefore consented to the storage of their data and that their data is made publicly available.

Balancing Risks and Benefits The suggestion of potential code reviewers is something that actual developers wish for and would benefit from, and it exists in practice in at least some limited fashion. For example, Henderson (2017) mentions that Google's code review system has the ability to suggest reviewers "*by looking at the ownership and authorship of the code being modified, the history of recent reviewers, and the number of pending code reviews for each potential reviewer.*"

There are clear risks to individuals if the generated reviewer profiles would be made public. Such profiles could be used to assess the individuals and could harm them professionally and/or publicly. However, the corresponding risk is low as reviewers and contributors are aware that their discussion is public and all reviews are expected to be done on a professional level and reviewers do not contribute sensitive information in their contributions. Developers contributing to the projects have a reasonable expectation that their contributions will be public as the terms and conditions of the three organisations make this clear.

The case study project limited the chance to reveal the generated profiles as much as possible, intentionally or inadvertently.

Current standards require making research as reproducible and replicable as much as possible. Usually that requires the release of prepared datasets in an open way. Data as generated by the code reviewer recommendation project is in conflict with such practices as such datasets should not contain data about individuals. To accommodate the need for open research, only the consolidated and accumulated data about precision, recall etc. per analysed software project can be stored and made available.

Fairness and Equity While the data available in the underlying repositories does not contain any data about race or gender, the names and email addresses of the contributors could point at such information. The project did explicitly not attempt to extract such information.

Compliance, Transparency and Accountability The investigation of the terms and conditions which was done for the CROP dataset case informed the considerations of the review recommendation project. However, it was necessary to investigate the changes to the terms and conditions since 2018 to ensure that such changes did not invalidate the arguments on which decisions were made. While there have been significant changes similar to the ones discussed for GitHub in Section 5, none of them affected the compliance of the project with the terms and conditions.

Summary The primary risks identified and addressed here were in large part the same as case study one. However, the requirement to maintain linkage across the datasets meant that the mitigations for privacy and reputation used in the first study were ineffective. This raised the risk level such that formal REC approval was then required and the arguments made across the other factors were needed to justify the work (e.g. arguing for implied consent/expectation).

7.4 Case 3: Using MSR Mining Challenges for Student Coursework

The third case study is not a dedicated research project. Instead, it describes how the MSR 2021 Mining Challenge was used to expose students to software engineering research in a coursework in which students have to come up with research ideas using the underlying dataset. The case study serves as an example of ethics considerations when using a third-party dataset.

The MSR Mining Challenges have been used as student coursework on the MSc Software Systems Engineering programme at UCL Computer Science. Students are asked to work in groups on the challenge in the same way one would work on an actual submission to the challenge and some groups have had work accepted (Bafatakis et al. 2019; Wilkie et al. 2018).

Each year, the potential ethics issues of the mining challenge have to be investigated. There are two challenges to the use of the mining challenge as a student coursework:

- The time available between the release of the MSR Mining Challenge and the start of the corresponding student coursework does not permit applying for approval via the university's REC. Instead, the student coursework needs to be in line with the UCL regulations for research projects that do not require ethics review.
- The structure and the time available for the coursework does not allow individual ethics consideration in line with UCL regulations per student group. Instead, the coursework needs to be shaped in a way that it allows the students to proceed with low risk of causing ethics conflicts.

For example, the MSR Mining Challenge of 2020 was based on the Software Heritage Archive, the ethics considerations for which (Software Heritage Archive 2019b) prevented its use as a student coursework in our context.

In this case study, we present how the 2021 MSR Mining Challenge (Allamanis et al. 2021) has been adapted to make it suitable for the described student coursework and address issues relating to consent and compliance.

The 2021 MSR Mining Challenge is based on the ManySStuBs4J dataset (Karampatsis and Sutton 2020), a dataset of 153,652 single statement bug fix changes mined from 1,000 popular open-source Java projects, annotated by whether they match any of a set of 16 bug templates.

As the dataset has been created from publicly available data without profiling of individuals, the dataset on its own does not pose any significant ethics issues. The dataset itself only contains data about code changes, their location, and the commit ids of the changes. It contains no data about individuals except for commit ids.

However, the conditions of using the dataset for analysis by students without the need for ethics review (under UCL's exemption criteria) required the creation of a subset of the dataset to contain data only about projects with a declared licence that is compatible with mining and analysing the project's repositories. In addition, the coursework explicitly disallowed to link the created dataset with any other dataset.

Identification of Stakeholders and Informed Consent As there is no strict possibility to limit the type of mining and analyses students would come up with, it was necessary to consider the need for informed consents differently where the contributors would allow any type of analysis done to their source code and repository. One way to assess this is to identify the licence of the repository and whether the licence is broad enough to cover this case.

GitHub allows projects to adopt a licence, which is then visible on the project's GitHub page. For each project in the ManySStuBs4J dataset, we retrieved the licence information via the GitHub API. The ManySStuBs4J dataset contains data from 634 GitHub projects, 540 of them have a declared licence. 104 have a licence that GitHub cannot automatically identify (but this does not necessarily mean that they are not licenced). Only the Apache License 2.0 (326), the MIT License (53), and the GNU General Public License v3.0 (24) are the adopted licences for more than 10 projects in the ManySStuBs4J dataset. All three of the licences are permissive enough that they can be interpreted in a way that allows the requirement of individual informed consent to be considered as met by licence application itself. Therefore, the subset of the ManySStuBs4J dataset that was provided to the students only consisted of the items from projects which adopted one of the three licences.

Balancing Risks and Benefits By limiting the students to work with the subset of the ManySStuBs4J dataset, the coursework did not only reduce the risk to the analysed projects' contributors, but more importantly, it did not expose the students to risks of violating ethics principles or software licences.

However, a planned submission on the coursework was abandoned because of the limited timescale available to seek UCL approval for studying code without appropriate licences.

Fairness and Equity The ManySStuBs4J dataset itself does not contain any sensitive information and students are not allowed to use commit ids to retrieve metadata.

Compliance, Transparency and Accountability By extracting a subset that is limited to three API-visible licences from the ManySStuBs4J dataset, it was ensured that students were in compliance with GitHub's Terms of Service and the projects' licences.

Summary In this case, the primary risk related to compliance was mitigated by restricting analysis to a reduced dataset of projects where appropriate licences had been automatically identified via the GitHub API, thus balancing effort with risk (and without substantively diminishing the experience for students). Restricting the exercise to only the data available in the reduced dataset meant that students were working solely with explicitly-published data under explicit licence (thus accounting for consent issues and intellectual property and falling outside the normal scope of ethics review).

7.5 Case 4: The Survey

To illustrate issues involved in direct participation research, as a final case study we describe the preparation and administration of the survey that we report in this paper. This took place over two rounds of data collection.

Prior to the first round, we prepared our materials (recruitment advert, information sheet, and questionnaire) and submitted these for review to the departmental research ethics committee, along with a description of our proposed investigation, how we intended to address the ethics issues, and why we believed the study met one of the criterion (anonymous survey) for exemption from full REC review.

In summary, our review request covered:

- Recruitment methods: ensuring that all spaces in which the advert appeared were ones in which we were identified as researchers to avoid any deception.
- Potential participation from our own institution and how power relationships there would be managed: via anonymity and non-institutional recruitment advertising platforms.
- Privacy and reputational risks to participants: managed through fully anonymous survey design (both in terms of direct data captured, and consideration of the mosaic re-identification potential arising from the questions collectively).
- Privacy by design: closed-form questions do not permit any identifiable information to be captured, and the survey questions collectively would not be sufficient to enable identification of an individual.
- Participant consent: given the anonymous nature of the survey and data capture, the presence of the full information sheet at the start of the questionnaire itself, and several granular consent questions requiring affirmative answers before the main questionnaire was revealed, we deemed participation to be sufficient evidence of consent and thus did not require a separate consent form to be signed.
- Availability of the information sheet: downloadable and part of the questionnaire itself.
- The survey questions.
- The planned data acquisition and management plan, and platform: study data were collected and managed using the REDCap electronic data capture tools hosted at UCL (Harris et al. 2009, 2019).
- An estimate of the completion time for the survey.
- Reward for participants: none.
- Period of survey availability
- Post-survey data management and sharing between the researchers.
- Data protection considerations: the survey was fully anonymous so the GDPR was not engaged.

Following some rounds of revision in response to departmental ethics review, we were granted approval to proceed under the relevant exemption. Owing to changes beyond our control in the way in which our recruitment advert would be displayed, we had to seek further authorisation to amend our original proposal to cover these new approaches (given new risks) after the original approval was given (a small illustration of Markham's "ethic is method, method is ethic" approach (Markham 2006): as our methods had to change, our ethical risk management had to respond).

Our first round of recruitment elicited few responses and it was kindly suggested by one of the anonymous reviewers of the first draft of this paper that we might try again and target representative members of the MSR community. We considered this in the context of UCL's policy framework, concluding that whilst there would be value in doing this, undertaking targeted recruitment would raise new ethical issues and take the work outside the scope of exemption at UCL. Since we could not be sure of obtaining approval in the time available for revision, we opted to repeat the first-round protocol using a similar recruitment method as before. We double-checked our intended approach with the departmental committee and were given approval to proceed without further authorisation since such changes as we made were in keeping with the conditions of the original authorisation and did not change the risk level.

Identification of Stakeholders and Informed Consent Stakeholders in this case were the participants in the survey. Informed consent was achieved through a full Participant Information Leaflet (PIL), available to download and keep, and as the preamble to the survey. Consent forms were not used (following the principle that survey participation can be considered evidence of consent where the risk in the survey itself is low). However, a number of questions to indicate understanding and consent were used at the start of the survey to control whether the main question set was shown, thus providing integral evidence of understanding and consent.

Balancing Risks and Benefits The risks in this case were designed to be low, avoiding identifiability by design both at the question and survey levels. The potential benefit was high from a high response rate.

Fairness and Equity The survey was publicly advertised in relevant fora and was open to all.

Compliance, Transparency and Accountability The PIL contained full contact information for the investigators, no deception was employed in the research methods, and matters such as data protection were considered fully. The platform was chosen to maintain anonymity.

Summary This was a straightforward low-risk survey study and thus the ethics issues involved are fairly standard in nature. They were considered explicitly and arguments formed for the ethical nature of each aspect.

7.6 Summary

The four example case studies illustrate actual ethics considerations that would not usually be discussed in the projects' publications. Moreover, they potentially demonstrate how the ethics considerations can shape projects before or while they commence, or even when projects have finished.

While the use of open source licencing is often seen as only relevant for copyright compliance, the first three case studies show how software licences can inform ethics decisions about stakeholders and informed consent.

The understanding of research ethics is evolving and informed by changes in legislation, terms and conditions of platforms like GitHub, and deep reflection on, and discussion of, the nature of the data and its provenance. Moreover, the four presented case studies have influenced not only the authors' view of research ethics in the area of Mining Software Repositories, but the necessary discussions with members of the research ethics panel led to a better understanding (for all) on how best to consider and treat some types of data (e.g. VCS metadata associated with commits; see the discussion in Section 5). This shows the importance of engaging with ethics approval bodies as part of MSR research activity and illustrates the general importance of dialogue in addressing research ethics.

8 Discussion

Researchers using data previously collected or assembled by others face a dilemma: Should they trust that the data has been collected or assembled for the purpose of their intended use in an ethical and lawful way, or do they need to try and verify this for themselves? The responsibilities of a researcher to their research participants are the same whether or not they are acquiring data from them directly or via a third-party who has previously collected the data. Not trusting that the data has been collected in an ethical way would prevent its usage (every data use would have to be preceded by a direct collection exercise), 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). The variety of ethics contexts and policies globally make this more difficult (e.g. one institution's view of ethically-sound research may differ from another's). Legal issues may also arise owing to different data protection regimes in force at the collecting and using sites, in some cases potentially inhibiting the use of data.

Even assuming that the data was collected in an ethical and lawful way originally does not mean that the intended research using the 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 a challenge or other research) requires re-consideration of ethics issues. 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 discussed in Section 7.2 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 page limits may act as a disincentive to including such sections in the face of reporting the primary results of research.

One might argue that simply adding a section to papers does not go far enough or recognise the extent to which ethics permeates investigations. The AoIR code (Franzke et al. 2020) highlights Markham's position that ethics is method and method is ethics (Markham 2006); in other words, that decisions made about research methods are inherently bound up in decisions about ethics and vice versa, and that this is an ongoing process of self-interrogation and reflection throughout a piece of research. Taken in this context, one might consider discussing ethics at many points in a research paper, not just in a separated section, thereby acting to *"counter a common presumption of 'ethics' as something of a 'one-off' tick-box exercise that is primarily an obstacle to research."* (Franzke et al. 2020).

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, similar to many other journals, already has a policy that authors should include a section on "Compliance with Ethical Standard" (Empirical Software Engineering 2020). 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 (2002): "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.

In times when papers are asked to provide their data to allow replication, there is also a significant challenge to find ways to not expose identifiable data in the research artefacts or replication packages (unless permission has been given for this).

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 (Hand 2018). 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 (Oezbek 2008). 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 direct consent in the MSR scenario may be difficult because of the etiquette and terms governing the use of repository information like email addresses as described above, e.g. data misuse (Software Heritage Archive 2019b) or developer annoyance (Baltes and Diehl 2016).

Another possibility might include the development of a licence that developers could attach to their profile governing the use of their repository data (the Apache License 2.0 already makes this likely unnecessary, and the recent GitHub changes to terms may make this available for a wider class of licences). Alternatively, the Menlo Report (Dittrich and

Kenneally 2012) suggests it might make sense to argue for a consent waiver from an oversight body on the grounds of impracticality.

Alternative approaches to consent and ethics matters may be found in other areas of internet-mediated research. Tuikka et al. (2017) 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 (2016) 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 (2016) 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. (2017) 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. This reflects the recommendations of the AoIR ethics code (Franzke et al. 2020). Perhaps the key requirement is the need to recognise that the nature of research undertaken on software repositories can vary widely, sometimes being more technical, at other times more social, and thus the ethical issues and frameworks that apply may lie outside the engineering discipline.

9 Guidelines

Whilst a "checklist" of ethical issues might be a desirable outcome of the discussion in this paper, it is essentially impossible to produce anything comprehensive of this nature. Judging what can be considered ethically-defensible requires evaluating aspects of the legislative, political, contractual, social, institutional, and personal contexts surrounding a proposed piece of research, along with the particular data and methods to be used, and the way in which the results will be reported. It is a dynamic, temporally and socially situated judgement. To anticipate every possible combination of these factors for all foreseeable time in such a way that a formulaic determination could be made is impossible. Where standards exist for particular communities these can be helpful to provide a common living reference point, e.g. the strongly values-based NIME Principles and Code of Ethics (Morreale et al. 2020), or standard, e.g. the ethics supplements in the ACM Empirical Standards for Software Engineering Research (Ralph et al. 2021), that can inform how ethics issues are commonly addressed.

To address the perceived desire for guidance arising from the survey, and since a checklist is unachievable, we offer instead a (by definition, partial) set of guidelines and questions for researchers to aid in considering the ethics of their MSR research (with a particular focus on consent, compliance and privacy). There are effectively two main stages: information gathering, and risk management (essentially formulating the ethical defence). In the first stage, one is seeking information that can support a case for the ethical use of the data in the absence of direct consent (e.g. consent, provenance), any constraints that might impinge on that use, any relevant legislative conditions that must be addressed, and the extent of personally-identifying data involved. In practical terms, consider the following:

1. If using a repository directly, check the licence terms for the repository contents (if a licence is applied). Does the licence applied permit study and/or research (or disclaim any restriction on use)? Does it also explicitly licence the metadata accompanying the code?
2. If there is no licence, or the licence covers the code only, what do the hosting site terms and conditions say about the use of the site contents? Do they permit or bar the intended use explicitly or implicitly?
3. If using a pre-collected repository published by a third-party, what information can be found about the ethics issues that were considered and procedures followed when it was collected? What restrictions (if any) are there on the proposed use of the data?
4. Does the data contain personally identifying information, either directly, or in combination with other accessible sources (e.g. a code fragment found in a public repository and thus linked to the originator)?

The particular arguments that apply in specific circumstances will vary and may be developed from the information determined in the above steps. The following three guidelines may help in formulating them in some common MSR situations. The way in which such arguments are then deployed will depend on the oversight bodies that govern the work to be undertaken and the processes that researchers need to follow (if any). In each case, the argument needs to address informed consent and compliance with relevant law (e.g. contract, copyright, and data protection). Other ethical matters such as risk to direct or indirect participants through the intended analysis, or privacy breaches as a result of research reporting are more directly in the control of the researchers and whilst they need to be addressed, do not pose such a difficult challenge to argue.

- If direct consent to study code is not being sought (as is likely in many MSR scenarios), argument for implicit ethical consent is needed. For licenced code, appeal to the licence terms to justify studying it. For unlicenced code, consider whether the disposition and/or prior use by others of the code offers support (e.g. have there been many downloads? has the code existed accessibly for a long time? has it been used in other studies?). An argument based on explicit licence terms is strong, one based on previous practice or technical availability less so but may be acceptable to an oversight body if it can be shown that harm is minimal or unlikely. The underlying principle is to demonstrate as far as possible that the intended use of the data is within the realistic expectation of those who provided it. Where a licence was attached, that offers a positive affirmation of intent. Without a licence, more assumptions are required about the intent (and thus expectation) of those who have donated code. In the absence of licence terms, consider also aspects of contract and copyright law as it applies in the jurisdiction of the research, and what risk would be taken by the research team and/or their organisation in proceeding without copyright permission, or in contravention of the site terms and conditions. Bear in mind data protection legislation requirements (e.g. a privacy notice to cover personal data contained within the code).
- If direct consent to study metadata is not being sought (as is likely in many MSR scenarios), argument for implicit ethical consent is again needed (for the same expectation-related reasons as above). Unless repository data is explicitly licensed (e.g. under the Apache License 2.0), consider the clarity of the terms offered to users both

in the way they are written and how clearly they are communicated when signing up. If these are clear, it might be argued that users reasonably expect their data to be used for research and thus consent is informed but implicit. Site terms and conditions may apply also here since they may constrain the use of such data (and thus contract law risks may exist). Data protection compliance aspects are likely to be relevant here and need actions to resolve. Similar arguments might apply to the study of communications (whether through issue trackers, mail archives, or forums).

- Where people will be directly involved in the research, standard human-studies factors will likely be involved. In other words, direct a-priori informed consent, free withdrawal, full information about the disposition of data and so forth will be needed. If recruiting using information found in software or metadata, consider site restrictions on such use of the data.

In addition to the guidelines above which may provide guidance through the typical ethics questions that need to be considered, we suggest some practices that are based on our experience in considering potential ethics issues and reducing risks.

- When extracting data for analysis, remove as much personal data as early as possible and anonymise data that cannot be removed.
- Do not mention names or user identifiers in papers.
- Do not use use email addresses found in MSR data for mass emails like questionnaires or asking for consent.
- Do not include projects without a licence in your MSR research unless it is absolutely necessary.
- When creating a dataset for public use, make potential users aware of the licences of the data from which the dataset has been created and whether there is any personal data in the dataset.
- Consider if your dataset could be used in a harmful way before releasing.
- State if your MSR research has received ethics approval or has been exempt.

The list above are good practices which cannot always be followed. However, if that is the case, the implications need to considered. Is the increased risk balanced with an increased benefit and are the burdens and benefits equitably distributed?

10 Threats

10.1 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 2006 to 2021, 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. Moreover, other datasets often are variations of the discussed datasets. For example, software distribution repositories such as Linux distributions or package repositories such as the npm Registry or Maven Central curate and publish software packages, and similar ethics considerations as discussed in Section 5.1 apply.

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 of website content or terms and conditions etc. in this paper was based on the versions accessible at the time of writing. 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. Such terms may also change in the future. 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, and future updates to such terms may affect the arguments that can be made. Careful scrutiny at the time of research is thus important.

In respect of our community survey, the primary threat arises from external validity owing to the very low number of responses and inability to generalise. We have mitigated this by not attempting to treat the data as anything more than anecdotal and indicative, and acknowledging its potential bias. We note also that the survey question designed to filter out bot responses (asking participants to pick a specific response) was answered incorrectly by three respondents. Since the other responses for those respondents were extremely coherent, we judged that this was likely an oversight by the respondents rather than a bot, and thus elected to include their answers in our discussion.

10.2 Ethics Considerations

In the spirit of following our own advice, we now discuss the ethics issues we considered in relation to the literature study and case studies. The community survey is described above as Case Study 4.

Literature 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 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. 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 presence or absence 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. In addition, to claim that something is “unethical” would require a universal definition of what constitutes “ethical” and as we have made clear elsewhere in this paper, there are many dynamic and contextual factors involved in deciding upon an ethically-defensible course of action. Thus in this paper, an absence of ethics discussion in the work we reviewed does not equate to an absence of ethics consideration: it is simply an absence of discussion.

Survey The survey study presented in Section 3 was reviewed by our departmental ethics committee (of which Gold is a member but was involved only as an applicant for approval in respect of this study), and then approved by the Head of Department in Computer Science in accordance with UCL policy for low-risk anonymous surveys. Approval for amendments was received directly from the Head of Department. Head of Department approvals do not generate a formal study number that can be quoted.

Case Studies The discussion of the four case studies is influenced by our understanding of the legal situation in the UK and the research ethics regulations currently in place at University College London (UCL) (UCL Research Ethics Committee 2020). As regulations and laws will differ at other organisations (and in other countries), the ethics issues around the projects discussed in the case studies could be considered differently at other organisations.

Guidelines The presented guidelines and good practices are based on our own experience of considering ethics, mainly in the context of the situation at UCL. They cannot be comprehensive across all times and places, and MSR researchers should not assume that further ethics considerations are not necessary if they follow the guidelines and practices.

11 Related Work

Professional codes of ethics, e.g. by IEEE-CS/ACM (Gotterbarn 2001), BCS (BCS: The Chartered Institute for IT 2021), or ACM (Association for Computing Machinery 2018), do not typically address research ethics directly (although the ACM code (Association for Computing Machinery 2018) 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 (Hand 2018). This reflects the ethical pluralism adopted by the AoIR (Franzke et al. 2020). Ethics has long been an integral part of research in most disciplines, including computing (Stahl et al. 2016) and software engineering with human studies in particular, see Hall and Flynn (2001), Singer and Vinson (2002) and Vinson and Singer (2008). 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 (Dittrich and Kenneally 2012).

At the same time as the 2001 Software engineering code of ethics and professional practice (Gotterbarn 2001) was developed, the focus turned to ethics issues in empirical software engineering, summarised by a special issue on research ethics for empirical software engineering (Singer and Vinson 2001). In the same issue, Hall and Flynn (2001) present survey results collected from 44 computer science departments in UK universities and highlight a number of issues.

Singer and Vinson (2002) 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 (Vinson and Singer 2001). Vinson and Singer (2008) extend their earlier work (Singer and Vinson 2002) 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 (2001) discuss ethics issues in such contexts.

El-Emam (2001) also raises a series of questions about the ethics implications of analysing open-source software, namely informed consent, minimisation of harm and confidentiality. German (2004) discusses the analysis of CVS repositories and raises multiple questions about ethics issues in such research, but does not answer them.

Robles et al. (2014) present the results of an online survey of 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 (2016) 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 (2017) 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.

Gonzalez-Barahona (2020) presented a tutorial at MSR 2020 on the implications of the GDPR for repository mining, highlighting the legal difficulties that may be encountered when doing so and offering potential routes for managing these.

12 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. Particular problems in MSR research (as in much internet-mediated research) are considerations for informed consent, expectation, risks to the data subjects, and compliance.

We presented the results of analysing MSR papers for the frequency of ethics discussion, and supplemented these with a small survey of researchers. These exercises showed that ethics discussion is not widespread in the MSR literature but that there may be value in developing guidance, not just to the community but to those with whom it interacts to seek authorisation for its research.

We thus 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 identified typical concerns and discussed their implications, finally drawing these together in some practical guidelines to support researchers in developing ethical defence for their work.

In summary, we argue that MSR researchers should consider their work from a variety of ethical positions to ascertain which frameworks may best apply and thereby find stronger ethical guidance and defence for their work (in turn improving the quality of research). We also raise the possibility that ethical discussion should permeate the presentation of research rather than being held as a separate and somewhat disconnected process.

Acknowledgements 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, Neil Oxtoby, Simon Parkin, Ross Purves, and Caroline Wardle.

Declarations No funding was received for conducting this work. In accordance with institutional policy as it applied at the relevant time, approvals were obtained as follows: Case Study 2 (Section 7.3) was approved by the UCL Research Ethics Committee (6082/002) and the survey reported in Section 7.5 was approved by the Head of Department of Computer Science, UCL. All other research presented in this work did not require approval, but was performed following advice from the appropriate sources at UCL at the relevant time.

See Section 10.2 for detailed discussion of the ethics considerations for the work presented in the paper.

References

Abric D, Clark OE, Caminiti M, Gallaba K, Mcintosh S (2019) Can duplicate questions on Stack Overflow benefit the software development community? In: 2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR). IEEE, pp 230–234. https://doi.org/10.1109/MSR.2019.00046

Allamanis M, Karampatsis RM, Sutton C (2021) Mining challenge. https://2021.msrconf.org/track/msr-2021-Mining-Challenge

An L, Mlouki O, Khomh F, Go A (2017) Stack Overflow: a code laundering platform? In: International Conference on Software Analysis, Evolution and Reengineering (SANER), pp 283–293

Andrews AA, Pradhan AS (2001) Ethical issues in empirical software engineering: the limits of policy. Empir Softw Eng 6(2):105–110. https://doi.org/10.1023/A:1011442319273

Association for Computing Machinery (2018) ACM code of ethics and professional conduct. https://doi.org/10.1145/3274591

Atwood J (2009) Stack Overflow creative commons data dump. https://stackoverflow.blog/2009/06/04/stack-overflow-creative-commons-data-dump/

Bacchelli A (2013) Mining challenge 2013: Stack Overflow. http://2013.msrconf.org/challenge.php

Badampudi D (2017) Reporting ethics considerations in software engineering publications. In: Proceedings of the 2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), pp 205–210. https://doi.org/10.1109/ESEM.2017.32

Badampudi D, Britto R, Unterkalmsteiner M (2019) Modern code reviews – preliminary results of a systematic mapping study. In: Proceedings of the Evaluation and Assessment on Software Engineering. ACM, New York, pp 340–345. https://doi.org/10.1145/3319008.3319354

Bafatakis N, Boecker N, Boon W, Cabello Salazar M, Krinke J, Oznacar G, White R (2019) Python coding style compliance on Stack Overflow. In: 2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR). IEEE, pp 210–214. https://doi.org/10.1109/MSR.2019.00042

Bailey M, Dittrich D, Kenneally E, Maughan D (2012) The Menlo report. IEEE Sec Priv 10(2):71–75. https://doi.org/10.1109/MSP.2012.52

Baltes S (2019) Software developers' work habits and expertise. PhD thesis, Universität Trier

Baltes S (2020) The SOTorrent Dataset. https://empirical-software.engineering/projects/sotorrent/

Baltes S, Diehl S (2016) Worse than spam. In: Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM). https://doi.org/10.1145/2961111.2962628

Baltes S, Diehl S (2019) Usage and attribution of Stack Overflow code snippets in GitHub projects. Empir Softw Eng 24(3):1259–1295. https://doi.org/10.1007/s10664-018-9650-5

Baltes S, Dumani L, Treude C, Diehl S (2018) SOTorrent. In: Proceedings of the 15th international conference on mining software repositories (MSR), pp 319–330. https://doi.org/10.1145/3196398.3196430

Baltes S, Treude C, Diehl S (2019) SOTorrent: Studying the origin, evolution, and usage of Stack Overflow code snippets. In: IEEE/ACM 16th International Conference on Mining Software Repositories (MSR), pp 191–194. https://doi.org/10.1109/MSR.2019.00038

Baysal O (2014) Mining challenge. http://2014.msrconf.org/challenge.php

BCS: The Chartered Institute for IT (2021) Code of conduct for BCS members. https://cdn.bcs.org/bcs-org-media/2211/bcs-code-of-conduct.pdf (last accessed 1st July 2021)

Beller M, Gousios G, Zaidman A (2017a) Mining challenge. http://2017.msrconf.org/#/challenge

Beller M, Gousios G, Zaidman A (2017b) TravisTorrent: Synthesizing Travis CI and GitHub for full-stack research on continuous integration. In: IEEE/ACM 14th International Conference on Mining Software Repositories (MSR), pp 447–450. https://doi.org/10.1109/MSR.2017.24

Benbunan-Fich R (2017) The ethics of online research with unsuspecting users: from a/b testing to c/d experimentation. Res Ethics 13(3-4):200–218. https://doi.org/10.1177/1747016116680664

Berry DM (2004) Internet research: privacy, ethics and alienation: an open source approach. Internet Res 14(4):323–332. https://doi.org/10.1108/10662240410555333

Bird C, Inoue K, Godfrey MW, Whitehead J (2009) MSR mining challenge 2009. http://2009.msrconf.org/challenge/

Bosu A, Sultana KZ (2019) Diversity and inclusion in open source software (OSS) projects: Where do we stand? In: 2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), IEEE, pp 1–11. https://doi.org/10.1109/ESEM.2019.8870179

British Educational Research Association (2018) Ethical guidelines for educational research, 4th edn. https://www.bera.ac.uk/publication/ethical-guidelines-for-educational-research-2018 (last accessed 1st July 2021)

Broad E, Smith A, Wells P (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 1st July 2021)

California State Legislature (2018) Assembly bill no. 375 – California consumer privacy act

Cortázar DI, Huesman N, Price A, de la Cruz VM (2018) Gender diversity analysis in the OpenStack community. https://superuser.openstack.org/wp-content/uploads/2018/06/Gender-Diversity-Analysis-in-the-OpenStack-Community-2018.pdf

Craver N (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

Creative Commons (2009) CC0 1.0 universal. https://creativecommons.org/publicdomain/zero/1.0/legalcode

Debian (2020a) Privacy policy. https://www.debian.org/legal/privacy

Debian (2020b) Ultimate Debian database. https://wiki.debian.org/UltimateDebianDatabase/

Dench S, Iphofen R, Huws U (2004) An EU code of ethics for socio-economic research. The Institute for Employment Studies, UK. http://www.respectproject.org/ethics/412ethics.pdf

Diehl S, Baltes S, Treude C (2019) Mining challenge. https://2019.msrconf.org/track/msr-2019-Mining-Challenge

Dittrich D, Kenneally E (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 1st July 2021)

Dittrich D, Kenneally E (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 1st July 2021)

Dyer R (2013) Bringing ultra-large-scale software repository mining to the masses with boa. PhD thesis, Iowa State University

Dyer R, Nguyen HA, Rajan H, Nguyen TN (2013) Boa: A language and infrastructure for analyzing ultra-large-scale software repositories. https://doi.org/10.1109/ICSE.2013.6606588

Dyer R, Nguyen HA, Rajan H, Nguyen TN (2015) Boa: Ultra-large-scale software repository and source-code mining. ACM Trans Softw Eng Methodol 25(1). https://doi.org/10.1145/2803171

Eclipse Foundation (2017a) Eclipse foundation software user agreement. https://www.eclipse.org/legal/epl/notice.php

Eclipse Foundation (2017b) Eclipse Public License – v 2.0. https://www.eclipse.org/legal/epl-2.0/

Eclipse Foundation (2019) Eclipse.org Terms of Use. https://www.eclipse.org/legal/termsofuse.php

El-Emam K (2001) Ethics and open source. Empir Softw Eng 6(4):291–292. https://doi.org/10.1023/A:1011962213685

Empirical Software Engineering (2020) Compliance with ethical standards https://www.springer.com/journal/10664/submissionguidelines#Instruction%20for%20Authors_Compliance%20with%20Ethical%20Standards

Franzke AS, Bechmann A, Zimmer M, Ess C (2020) Internet research: Ethical Guidelines 3.0 https://aoir.org/reports/ethics3.pdf (last accessed 1st July 2021)

Free Software Foundation (2021) Various licenses and comments about them. https://www.gnu.org/licenses/license-list.en.html (last accessed 1st July 2021)

Fry T, Dey T, Karnauch A, Mockus A (2020) A dataset and an approach for identity resolution of 38 million author ids extracted from 2B Git commits. In: Proceedings of the 17th International Conference on Mining Software Repositories. ACM, New York, pp 518–522. https://doi.org/10.1145/3379597.3387500

German D (2004) Mining CVS repositories, the softChange experience. In: International workshop on mining software repositories (MSR), pp 17–21. https://doi.org/10.1049/ic:20040469

GitHub Inc (2020a) GitHub terms of service: D. user-generated content. https://docs.github.com/en/free-pro-team@latest/github/site-policy/github-terms-of-service#d-user-generated-content

GitHub Inc (2020b) Updates to our terms of service and privacy statement. https://github.blog/2020-10-15-updates-to-our-terms-of-service-and-our-privacy-statement/ (last accessed 1st July 2021)

GitHub Inc (2021) Information usage restrictions. https://docs.github.com/en/github/site-policy/github-acceptable-use-policies#5-information-usage-restrictions/ (last accessed 8th July 2021)

Gold NE, Krinke J (2020a) Ethical mining – a case study on MSR mining challenges. https://www.youtube.com/watch?v=wYz7DLLJa-c

Gold NE, Krinke J (2020b) Ethical mining: A case study on MSR mining challenges. In: Proceedings of the 17th International Conference on Mining Software Repositories. ACM, New York, pp 265–276. https://doi.org/10.1145/3379597.3387462

Gonzalez-Barahona JM (2020) Mining software repositories while respecting privacy. https://2020.msrconf.org/details/msr-2020-Education/1/Mining-Software-Repositories-While-Respecting-Privacy

Gonzalez-Barahona JM, Robles G, Izquierdo-Cortazar D (2015) The MetricsGrimoire database collection. In: 2015 IEEE/ACM 12th Working Conference on Mining Software Repositories. IEEE, pp 478–481. https://doi.org/10.1109/MSR.2015.68

Gotterbarn D (2001) Software engineering code of ethics and professional practice. Sci Eng Ethics 7(2):231–238. https://doi.org/10.1007/s11948-001-0044-4

Gousios G (2013) The GHTorent dataset and tool suite. In: 10th Working Conference on Mining Software Repositories (MSR), pp 233–236. https://doi.org/10.1109/MSR.2013.6624034

Hall T, Flynn V (2001) Ethical issues in software engineering research: a survey of current practice. Empir Softw Eng 6(4):305–317. https://doi.org/10.1023/A:1011922615502

Han D, Ragkhitwetsagul C, Krinke J, Paixão M, Rosa G (2020) Does code review really remove coding convention violations? In: International Working Conference on Source Code Analysis and Manipulation (SCAM), pp 43–53

Hand DJ (2018) Aspects of data ethics in a changing world: Where are we now? Big Data 6(3). https://doi.org/10.1089/big.2018.0083

Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG (2009) Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 42(2):377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, McLeod L, Delacqua G, Delacqua F, Kirby J, Duda SN (2019) The REDCap consortium: Building an international community of software platform partners. J Biomed Inform 95:103208. https://doi.org/10.1016/j.jbi.2019.103208

Henderson F (2017) Software engineering at Google. arXiv:1702.01715 [cs.SE]

Hindle A (2010) MSR mining challenge 2010. http://2010.msrconf.org/challenge/

Hindle A, Herraiz I, Shihab E, Jiang ZM (2010) Mining challenge 2010: Freebsd, gnome desktop and debian/ubuntu. In: 7th IEEE Working Conference on Mining Software Repositories (MSR), pp 82–85. https://doi.org/10.1109/MSR.2010.5463350

Horwitz J (2020) Facebook fights research project. Wall Street Journal, Eastern edition

Karampatsis RM, Sutton C (2020) How often do single-statement bugs occur? In: Proceedings of the 17th International Conference on Mining Software Repositories. ACM, New York, pp 573–577. https://doi.org/10.1145/3379597.3387491

Kim S, Hassan AE, Lanza M, Godfrey MW (2008) MSR mining challenge 2008. http://2008.msrconf.org/challenge/

Kshetri N, Voas J (2020) Thoughts on general data protection regulation and online human surveillance. Computer 53(1):86–90. https://doi.org/10.1109/MC.2019.2951984

Lin B, Serebrenik A (2016) Recognizing gender of stack overflow users. In: Proceedings of the 13th International Conference on Mining Software Repositories. ACM, pp 425–429. https://doi.org/10.1145/2901739.2901777

Locatelli E (2020) Academy/industry partnership and corporate data: ethical considerations. IRE 3.0 Companion 6.2. https://aoir.org/reports/ethics3.pdf, (last accessed 1st July 2021)

Markham A (2006) Ethic as method, method as ethic: a case for reflexivity in qualitative ICT research. J Inf Ethics 15(2):37–54. https://doi.org/10.3172/JIE.15.2.37

Markham A, Buchanan E (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 1st July 2021)

Markovtsev V, Long W (2018) Public Git archive. In: Proceedings of the 15th International Conference on Mining Software Repositories – MSR '18. ACM Press, New York, pp 34–37. https://doi.org/10.1145/3196398.3196464

Matalonga H, Cabral B, Castor F, Couto M, Pereira R, de Sousa SM, Fernandes JP (2019) GreenHub farmer: Real-world data for Android energy mining. In: 2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR). IEEE, pp 171–175. https://doi.org/10.1109/MSR.2019.00034

Menzies T, Bird C, Zimmermann T (2014) Analyzing software data: after the gold rush (a goldfish-bowl panel). In: Companion Proceedings of the 36th International Conference on Software Engineering – ICSE Companion 2014. ACM Press, New York, pp 103–104. https://doi.org/10.1145/2591062.2594395

Miller FG, Rosenstein DL (2002) Reporting of ethical issues in publications of medical research. The Lancet 360(9342):1326–1328. https://doi.org/10.1016/S0140-6736(02)11346-8

Morreale F, Gold N, Chevalier C, Masu R (2020) NIME principles & code of practice on ethical research. https://www.nime.org/ethics/

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 1st July 2021)

Naughton J (2020) Facebook has good reasons for blocking research into political ad targeting. The Guardian https://www.theguardian.com/commentisfree/2020/oct/31/facebook-has-good-reasons-for-blocking-research-into-political-ad-targeting

Nguyen H, Dyer R (2016) Mining challenge. http://2016.msrconf.org/#/challenge

Oezbek C (2008) Research Ethics for studying open source projects. In: Proceedings of 4th Research Room FOSDEM http://www.inf.fu-berlin.de/inst/ag-se/pubs/OSSethics-2008.pdf (last accessed 1st July 2021)

Open Infrastructure Foundation (2018) Open development. https://opendev.org/osf/four-opens/src/doc/source/opendevelopment.rst

Open Source Guides (2019) The legal side of open source. https://opensource.guide/legal/ (last accessed 1st July 2021)

Open Source Initiative (2019) Licenses by name. https://opensource.org/licenses/alphabetical (last accessed 1st July 2021)

Paixão M, Krinke J, Han D, Ragkhitwetsagul C, Harman M (2017) Are developers aware of the architectural impact of their changes? In: International Conference on Automated Software Engineering (ASE), pp 95–105

Paixão M, Krinke J, Han D, Harman M (2018) CROP: Linking code reviews to source code changes. In: Proceedings of the 15th International Conference on Mining software Repositories, pp 46–49

Paixão M, Krinke J, Han D, Ragkhitwetsagul C, Harman M (2019) The impact of code review on architectural changes. IEEE Trans Softw Eng

Paixão M, Uchôa A, Bibiano AC, Oliveira D, Garcia A, Krinke J, Arvonio E (2020) Behind the intents: An in-depth empirical study on software refactoring in modern code review. In: International conference on mining software repositories (MSR), pp 125–136

Pietri A, Spinellis D, Zacchiroli S (2019) The software heritage graph dataset: Public software development under one roof. In: 2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR). IEEE, pp 138–142. https://doi.org/10.1109/MSR.2019.00030

Pietri A, Spinellis D, Zacchiroli S (2020a) Mining challenge. https://2020.msrconf.org/track/msr-2020-Mining-Challenge

Pietri A, Spinellis D, Zacchiroli S (2020b) The software heritage graph dataset. In: Proceedings of the 17th International Conference on Mining Software Repositories. ACM, New York, pp 1–5. https://doi.org/10.1145/3379597.3387510

Pinzger M, Gall H, Lanza M, D'Ambros M (2006) MSR mining challenge 2006. http://2006.msrconf.org/challenge/

Proksch S (2017) Enriched event streams: a general platform for empirical studies on in-IDE activities of software developers. PhD thesis, Technische Universitat Darmstadt

Proksch S (2019) KaVE Project. http://www.kave.cc/

Proksch S, Amann S, Nadi S (2018a) Enriched event streams. In: Proceedings of the 15th international conference on mining software repositories (MSR), pp 62–65. https://doi.org/10.1145/3196398.3196400

Proksch S, Amann S, Nadi S (2018b) Mining challenge. https://2018.msrconf.org/track/msr-2018-Mining-Challenge

Ragkhitwetsagul C, Krinke J, Paixão M, Bianco G, Oliveto R (2019) Toxic code snippets on stack overflow. IEEE Trans Softw Eng. https://doi.org/10.1109/TSE.2019.2900307

Ralph P, bin Ali N, Baltes S, Bianculli D, Diaz J, Dittrich Y, Ernst N, Felderer M, Feldt R, Filieri A, de França BBN, Furia CA, Gay G, Gold N, Graziotin D, He P, Hoda R, Juristo N, Kitchenham B, Lenarduzzi V, Martínez J, Melegati J, Mendez D, Menzies T, Molleri J, Pfahl D, Robbes R, Russo D, Saarimäki N, Sarro F, Taibi D, Siegmund J, Spinellis D, Staron M, Stol K, Storey MA, Taibi D, Tamburri D, Torchiano M, Treude C, Turhan B, Wang X, Vegas S (2021) Empirical standards for software engineering research. arXiv:2010.03525 [cs.SE]

Rao N, Bansal C, Guan J (2021) Search4Code: Code search intent classification using weak supervision. In: International Conference on Mining Software Repositories (MSR), pp 575–579. https://doi.org/10.1109/MSR52588.2021.00077

Robles G, Arjona Reina L, Serebrenik A, Vasilescu B, González-Barahona JM (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), pp 396–399. https://doi.org/10.1145/2597073.2597129

Schröter A (2011a) Mining challenge. http://2011.msrconf.org/msr-challenge.html

Schröter A (2011b) MSR challenge 2011. In: Proceeding of the 8th working conference on mining software repositories (MSR). https://doi.org/10.1145/1985441.1985478

Shihab E (2012) Mining challenge. http://2012.msrconf.org/challenge.php

Shihab E, Kamei Y, Bhattacharya P (2012) Mining challenge 2012: The Android platform. In: 9th IEEE Working Conference on Mining Software Repositories (MSR), pp 112–115. https://doi.org/10.1109/MSR.2012.6224307

Singer J, Vinson N (2001) Why and how research ethics matters to you. yes, you! Empir Softw Eng 6(4):287–290. https://doi.org/10.1023/A:1011998412776

Singer J, Vinson NG (2002) Ethical issues in empirical studies of software engineering. IEEE Trans Softw Eng 28(12):1171–1180. https://doi.org/10.1109/TSE.2002.1158289

Software Heritage Archive (2019a) Software heritage: ethical charter for mirrors. https://www.softwareheritage.org/legal/mirrors-ethical-charter/ (last accessed 1st July 2021)

Software Heritage Archive (2019b) Software Heritage: Ethical Charter for using the archive data. https://www.softwareheritage.org/legal/users-ethical-charter/ (last accessed 1st July 2021)

Software Heritage Archive (2019c) Software Heritage: Terms of use for bulk access. https://www.softwareheritage.org/legal/bulk-access-terms-of-use/ (last accessed 1st July 2021)

Soto-Valero C, Bourcier J, Baudry B (2018) Detection and analysis of behavioral T-patterns in debugging activities. In: Proceedings of the 15th international conference on mining software repositories. ACM, pp 110–113. https://doi.org/10.1145/3196398.3196452

Stahl BC, Timmermans J, Mittelstadt BD (2016) The ethics of computing: a survey of the Computing-Oriented literature. ACM Comput Surv 48(4). https://doi.org/10.1145/2871196

Sugiura L, Wiles R, Pope C (2017) Ethical challenges in online research: Public/private perceptions. Res Ethics 13(3-4):184–199. https://doi.org/10.1177/1747016116650720

The Apache Software Foundation (2004) Apache license version 2.0. https://www.apache.org/licenses/LICENSE-2.0.txt

The British Psychological Society (2014) Code of human research ethics

The British Psychological Society (2018) Code of ethics and conduct

The European Parliament and the Council of the European Union (2016) General data protection regulation (EU) 2016/679. Official Journal of the European Union https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679, (last accessed 1st July 2021)

The FreeBSD Project (2012) FreeBSD's Privacy Policy. https://www.freebsd.org/privacy.html

The Linux Foundation (2018) Summary of GDPR Concepts for Free and Open Source Software Projects. https://www.linuxfoundation.org/wp-content/uploads/lf_gdpr_052418.pdf (last accessed 1st July 2021)

The TestRoots Team (2020) TravisTorrent: Free and Open Travis Analytics for Everyone. https://travistorrent.testroots.org

Thomas DR, Pastrana S, Hutchings A, Clayton R, Beresford AR (2017) Ethical issues in research using datasets of illicit origin. In: Proceedings of the ACM SIGCOMM internet measurement conference (IMC). https://doi.org/10.1145/3131365.3131389

Townsend L, Wallace C (2016) Social media research: A guide to ethics https://www.gla.ac.uk/media/Media_487729_smxx.pdf (last accessed 1st July 2021)

Travis CI (2019) Privacy policy. https://docs.travis-ci.com/legal/privacy-policy/

Tuikka AM, Nguyen C, Kimppa KK (2017) Ethical questions related to using netnography as research method. ORBIT J 1(2). https://doi.org/10.29297/orbit.v1i2.50

UCL Research Ethics Committee (2020) Research Ethics at UCL. https://ethics.grad.ucl.ac.uk/

U.S. Department of Health and Human Services (1996) Health Insurance Portability and Accountability Act of 1996 P.L. No. 104-191

Vasilescu B, Posnett D, Ray B, van den Brand MG, Serebrenik A, Devanbu P, Filkov V (2015) Gender and tenure diversity in GitHub teams. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, pp 3789–3798. https://doi.org/10.1145/2702123.2702549

Vinson N, Singer J (2001) Getting to the source of ethical issues. Empir Softw Eng 6(4):293–297

Vinson NG, Singer J (2008) A practical guide to ethical research involving humans. In: Shull F, Singer J, DIK Sjøberg (eds). Springer, London, pp 229–256. https://doi.org/10.1007/978-1-84800-044-5

Whitney SN (2016) Balanced ethics review: a guide for institutional review board members. Springer International Publishing. https://doi.org/10.1007/978-3-319-20705-6

Wilkie J, Halabi ZA, Karaoglu A, Liao J, Ndungu G, Ragkhitwetsagul C, Paixão M, Krinke J (2018) Who's this? developer identification using IDE event data. In: Proceedings of the 15th International Conference on Mining Software Repositories – MSR '18. ACM Press, New York, pp 90–93. https://doi.org/10.1145/3196398.3196461

Yamashita A, Abtahizadeh SA, Khomh F, Gueheneuc YG (2017) Software evolution and quality data from controlled, multiple, industrial case studies, IEEE. https://doi.org/10.1109/MSR.2017.44

Yamashita A, Petrillo F, Khomh F, Guéhéneuc YG (2018) Developer interaction traces backed by IDE screen recordings from think aloud sessions. In: Proceedings of the 15th International Conference on Mining Software Repositories – MSR '18. ACM Press, New York, pp 50–53. https://doi.org/10.1145/3196398.3196457

Ying A (2015) Mining challenge. http://2015.msrconf.org/challenge.php

Zimmermann T (2007) MSR mining challenge 2007. http://2007.msrconf.org/challenge/

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Nicolas Gold is an Associate Professor of Computer Science at University College London in the UK. His current research interests include software engineering (in particular, source code analysis), ethics, and music computing (in particular, data sonification and musification in healthcare, and music and making in education). He is a Fellow of the British Computer Society.

Jens Krinke is an Associate Professor in the Software Systems Engineering Group at the University College London, where he is Director of CREST, the Centre for Research on Evolution, Search, and Testing. His main focus is software analysis for software engineering purposes. His current research interests include mining software repositories, software similarity and reuse, and modern code review. He is well known for his work on program slicing and clone detection.