How Should Software Engineering Secondary Studies Include Grey Material?

Barbara Kitchenham, *Member, IEEE,* and Lech Madeyski, *Senior Member, IEEE* and David Budgen, *Member, IEEE*

Abstract—*Context*: Recent papers have proposed the use of *grey literature* (GL) and multivocal reviews. These papers have raised issues about the practices used for systematic reviews (SRs) in software engineering (SE) and suggested that there should be changes to the current SR guidelines. *Objective*: To investigate whether current SR guidelines need to be changed to support GL and multivocal reviews. *Method*: We discuss the definitions of GL and the importance of GL and of industry-based field studies in SE SRs. We identify properties of SRs that constrain the material used in SRs: a) the nature of primary studies; b) the requirements of SRs to be auditable, traceable, and reproducible; and explain why these requirements restrict the use of blogs in SRs. *Results*: SR guidelines have always considered GL as a possible source of primary studies and have never supported exclusion of field studies that incorporate the practitioners' viewpoint. However, the concept of GL, which was meant to refer to documents that were not formally published, is now being extended to information from sources such as blogs/tweets/Q&A posts. Thus, it might seem that SRs do not make full use of GL because they do not include such information. However, the unit of analysis for an SR is the primary study. Thus, it is not the *source* but the *type* of information that is important. Any report describing a rigorous empirical evaluation is a candidate primary study. Whether it is actually included in an SR depends on the SR eligibility criteria. However, any study that cannot be guaranteed to be publicly available in the long term should not be used as a primary study in an SR. This does not prevent such information from being aggregated in surveys of social media and used in the context of evidence-based software engineering (EBSE). *Conclusions*: Current guidelines for SRs do not require extensions, but their scope needs to be better defined. SE researchers require guidelines for analysing social media posts (e.g., blogs, tweets, vlogs), but these should be based on qualitative primary (not secondary) study guidelines. SE researchers can use mixed-methods SRs and/or the fourth step of EBSE to incorporate findings from social media surveys with those from SRs and to develop industry-relevant recommendations.

Index Terms—evidence-based software engineering, systematic reviews, systematic mapping studies, mixed-methods reviews, grey literature, multivocal reviews

1 Introduction

The role that *grey literature* can or should fulfil in systematic reviews has recently become a subject of interest in software engineering research. For example, a special issue of *Information and Software Technology* [1] was based on grey literature and multivocal reviews. In addition, there are several mapping studies that have reviewed the use of grey literature in systematic reviews (see [2], [3], and [4]).

However, we believe that some issues related to the use of grey literature need better clarification to avoid misunderstandings about the nature of systematic reviews and their goals, methodology, and limitations. One issue of concern is that the term *grey literature* is not used consistently in software engineering, which can lead to misunderstandings about the scope of systematic reviews. For example, Zhang et al. [4] reviewed the definitions of grey literature implied by the eligibility criteria used in systematic reviews. They identified three possible definitions of grey literature. All three definitions emphasized the lack of peer-review as a feature of grey literature, but each definition also considered other criteria, including the accessibility of the material (i.e., public, not private), the origin of the material (i.e., organizational not individual), or the quality of the material (i.e., scientific rather than non-scientific). All of these definitions would identify books and book chapters as grey literature. However, the recommendation to include grey material reported in the systematic review guidelines used in software engineering (e.g., [5] and [6]) was based on the definitions of grey literature developed by librarians and information scientists, which identify books and book chapters as white literature, and lack of formal publication as the main criteria for defining grey literature.

Thus, our high level research question is whether current SR guidelines need to be revised to address grey literature? We do not have more detailed research questions but report instead our original views on the topic of grey literature reviews and multivocal reviews[1]. The views that are the basis of the topics we discuss in this article are:

1) Systematic Reviews can include grey literature, providing it conforms with the SR eligibility criteria.
2) Social media posts such as blogs and tweets can identify new solutions and new ideas, but do not usually report

- *B. Kitchenham is with the School of Computing and Mathematics, Keele University, Keele, Staffordshire, ST5 5BG, UK.ORCID: 0000-0002-6134-8460*
- *Lech Madeyski (✉) is with the Faculty of Computer Science and Management, Wroclaw University of Science and Technology, Wyb.Wyspianskiego 27, 50370 Wroclaw, Poland.*
  E-mail: Lech.Madeyski@pwr.edu.pl ORCID: 0000-0003-3907-3357
- *David Budgen is with the Department of Computer Science, Durham University, Durham DH1 3LE, UK. ORCID: 0000-0001-7143-0241*

Manuscript received April 1, 2022. (Corresponding author: Lech Madeyski)

1. Which are similar to the propositions that can be used to identify important issues in case study research [7].

the details of any empirical studies evaluating such ideas and solutions.

3) Blogs, tweets etc. are not the only source of industry-based and practitioner-based viewpoints. If available, reports of industry field studies should always be included, otherwise SRs only provide weak evidence.
4) Surveys of social media sources can be used to understand and interpret SR results in mixed-methods studies.

We discuss various ways of defining grey literature and the implications of these definitions in Section 2, where we examine the Prague definition of grey literature [8], which we suggest should be the definition adopted for systematic reviews. The Prague definition excludes internet information sources such as blogs, tweets, and e-mails. The adoption of different definitions of *grey literature* explains why we believe SR guidelines properly address grey literature while other researchers do not [9].

In Section 3, we identify primary studies as the *unit of analysis* for SRs, although we note that mapping studies tend to classify studies at the source level. Primary studies report evidence based on rigorous empirical investigations which are unlikely to be reported in social media posts. We also point out that systematic reviews and mapping studies both have requirements for audibility, traceability, and reproducibility that are unlikely to be met by literature that does not conform with the Prague definition of grey literature.

We regard evidence from primary studies as being critical for SRs, and provide examples of problems arising from *expert opinion*-based assessments found in medicine and software engineering. In Section 4, we discuss what sort of studies are needed to represent the viewpoint of SE practitioners. We point out that social media posts are not the only source of the practitioner viewpoint and, in particular, field studies reported in academic sources should not be ignored.

In addition, social media posts raise their own analysis problem as discussed in Section 5. In Section 6, we examine some of the ambiguities in the guidelines for multivocal reviews [9], in particular, whether the analysis of social media posts should be considered as a secondary study method or as a form of primary study, and the related issue of which form should be adopted by the guidelines for analysing social media posts.

In Section 7, we discuss how the findings from social media surveys can be used to support academic research, practice and decision making. In particular, we introduce the concept of mixed-methods reviews as a means of incorporating social media information into SRs that does not compromise the integrity of the empirical evidence. In addition, we suggest that social media surveys can contribute to the fourth step of the evidence-based software engineering process, which aims to integrate critical appraisal with software engineering expertise and stakeholders' values (see [6], [10]). We summarise our arguments and present our conclusions and recommendations in Section 8.

2 Defining Grey Literature

The term *grey literature* as used in systematic review guidelines refers to *unpublished primary studies* that conform to the *Luxembourg* definition:

> "information produced on all levels of government, academia, business and industry in electronic and print formats not controlled by commercial publishing, i.e., where publishing is not the primary activity of the producing body."

This definition was developed by librarians and information scientists and was agreed at the Third International Conference on Grey Literature in 1997[2]. It was subsequently expanded at the New York conference in 2004. Using this definition, academic and industry technical reports, government and industry white papers, and academic theses (PhD, MSc, or BSc) would be classified as grey literature.

The *New York* definition was further refined at the 12th International Conference on Grey Literature held in Prague in December 2010, where a new approach to grey literature was discussed [8]. The rationale for the changes was that, while the existing definition of grey literature remained helpful and should not be replaced, it needed to be adapted to the context of internet publishing, to consider issues such as protection of intellectual property and identification of a minimum quality level (by means of peer review, or other external validation). This led to the following *Prague* definition:

> "Grey literature stands for manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers i.e., where publishing is not the primary activity of the producing body."

This definition emphasises the *collection* and *preservation* of grey material and is consistent with the goal of librarians to catalogue and archive important information. As a definition, it is important for systematic reviews because it includes:

1) The type of non-white documents that are most likely to provide evidence derived from rigorous empirical studies (referred to as *primary studies* in the context of SRs). This is discussed further in Section 3.
2) The type of non-white documents that are most likely to remain accessible in the public domain *in the long term*. This addresses the goal of systematic reviews to be as auditable, traceable and reproducible as possible. This is discussed further in Section 5.

More recently, researchers in many empirical disciplines have proposed extending the concept of grey literature to include material arising from social media such as blogs, tweets, Q&A fora, videos, emails, etc.

R. Adams et al. [12] present a tiered-model which defines grey literature in terms of outlet control and credibility. Grey Literature tier 1 includes information sources with high outlet control and high credibility, such as books, magazines, government reports, and white papers. Grey Literature tier 2 includes information sources with moderate outlet control and moderate credibility, such as annual reports, news articles, presentations, videos, Q&A sites (such as

2. For more information in the history of grey literature see Rucinski [11].

StackOverflow), and Wiki articles. Grey Literature tier 3 includes information sources with low outlet control and low credibility and includes blogs, emails, and tweets.

J. Adams et al. [13] suggest another model based on three types of information source: *grey literature* such as internal reports, working papers and newsletters which they classify as *informally published*; *grey data* such as tweets, blogs, Facebook status updates, which they classify as *self-published*; and *grey information* such as meeting notes, personal e-mails, and personal memories which they classify as *unpublished*. J. Adams' model provides different names for different types of information but does not explicitly mention academic theses as a form of grey literature.

These models are not exactly equivalent and are largely defined by example, which is a weak method of definition, because such definitions are seldom complete, and as new examples occur, they may be classified differently by different people. An issue stemming from this is that the term *grey literature* clearly means very different things to different people, so that any discussion of grey literature needs to be very specific about the scope of the discussion.

In the context of software engineering, Garousi et al. [9] discussed the model proposed by R. Adams et al., but, in [1], Garousi et al. proposed that software engineering researchers adopt the following definition of grey literature:

> "Grey literature in SE can be defined as any material about SE that is not formally peer-reviewed nor formally published."

In [9] Garousi et al. criticise systematic reviews, saying:

> "Unfortunately, SLRs fall short in providing full benefits since they typically review the formally-published literature only while excluding the large bodies of the "grey" literature (GL)."

It is clear that social media sources represent large bodies of material, but, apart from grey literature conforming with the Prague definition, we do not agree that other informally or self-published material is appropriate for inclusion in SRs.

To emphasize the difference between grey literature and other non-white sources, we prefer to use the term grey literature to refer *only* to information of the form defined by the Prague definition, from which the following information sources are of particular relevance to systematic reviews:

- PhD and Masters theses,
- academic technical reports,
- industry and government white papers,
- versions of papers, and their supplementary materials that are *in press*, or published on pre-print, archive, or protocol registration sites.

It is important to appreciate that guidelines for systematic reviews have always permitted the use of grey literature, as defined originally by the Luxembourg definition and now by the Prague definition, as a source of primary studies for systematic reviews.

We advocate that information from other informal sources should be defined in terms that describe the information source accurately and avoid any overlap with the Prague definition of grey literature. Based on the discussion in [13], we propose to make the distinction clear by using the following two terms:

1) *Social media posts*, when referring to online communication media such as blogs, tweets, wiki's, vlogs, online videos, Q&A fora. Furthermore, although we are using social media as a generic term, we strongly recommend using more specific terms when talking about different types of online material. The problem with information obtained from these types of document is that, although the material may have been easily accessible at a specific point in time, it is not guaranteed to have *long-term accessibility in the public domain*.
2) *Personal communications*, when referring to industry and government internal communications such as memos, e-mails, meeting notes, minutes and agendas. The problem with information obtained from sources such as these is that they are not publicly accessible, which is related to the aforementioned lack of long-term accessibility in the public domain.

For completeness, we refer to conventionally published and catalogued information sources such as books and book chapters as *white literature*, although such information sources are sometimes explicitly excluded from SE systematic reviews, and would be classified as grey literature by some SE researchers. The Luxembourg and Prague definitions clearly exclude such publications from their definition of grey literature. Our current position is that any conventionally published book and any of its individual chapters is the same as any other conventionally published material and should be classified as *white* literature. Figure 1 summarises the terminology we use in this article.

Referring to all types of unpublished material as *grey literature*, makes it seem appropriate to treat all unpublished information as appropriate for inclusion in systematic reviews. However, social media posts of different types contain very different types of information from grey literature conforming with the Prague definition. They need to be analysed in ways that reflect the specific type of information they provide. Furthermore, to be found by interested readers, study reports (whether white or grey) need titles that specify the type of online information that they are investigating (see, for example, any of the numerous reports of studies of open source software or StackOverflow published in both white and grey literature sources).

In Section 3 we point out that the unit of analysis is the primary study and that, apart from formally published material, only grey literature according to the Prague definition is likely to include reports of primary studies.

In Section 7, we discuss the use of social media posts and personal communications in software engineering research.

3 Admissible Primary Studies

Systematic reviews analyse the findings from primary studies, where these report the outcomes of *completed research projects*. From this viewpoint, it is not the *source* of the primary study that is important, it is the *nature* of the primary study itself. Furthermore, systematic reviews aiming to influence practice need *evidence* derived from independent studies (quantitative or qualitative) that have used a defined empirical research method. In the context of SRs, searching for grey literature, as defined by the Prague definition, is considered to be important to avoid the publication bias that

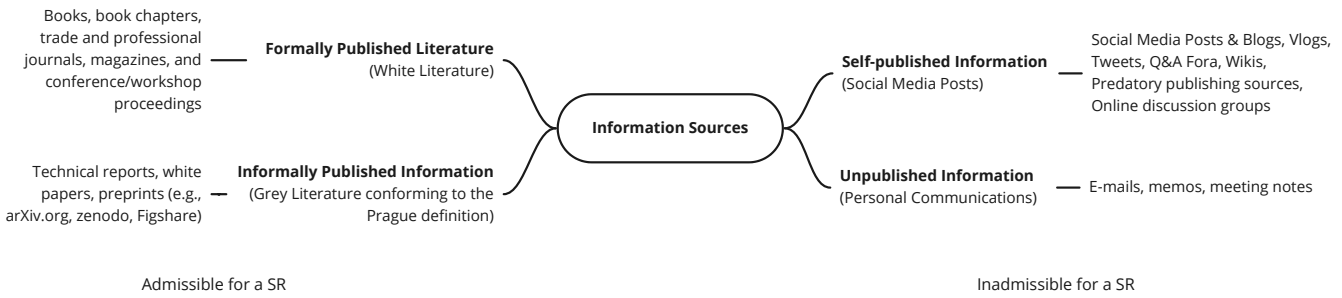

Fig. 1. Types of Information Source

may occur where articles describe primary studies that did not find novel results. Such articles may not be formally published, because authors, reviewers, or journal editors are not very interested in replications or negative results.

Grey literature, as defined by the Prague definition, which includes preprints, PhD theses, technical reports, white papers, etc. often includes reports of primary studies that have negative or inconclusive results. Therefore, to avoid bias, any such studies should be considered as candidate primary studies in an SR, and included as primary studies, providing that they conform to the SR eligibility criteria. No special guidelines are needed for processing such primary studies, because after passing the eligibility criteria they are regarded as equivalent to primary studies from white literature sources.

Recent reviews of the use of grey literature have confirmed that SE systematic reviews do include grey literature. For example, Kamei et al. [2] found 126 SRs out of a total of 446 that included references to grey literature. Unfortunately, they included book and book chapters in their definition of grey literature, as well as social media sources. In terms of grey literature conforming to the Prague definition, they found 53 references to technical reports, 34 to theses, 11 to white papers, and 5 to preprints. In addition, sources in other categories they used might have been covered by the Prague definition, in particular, web documents (8 references) and magazine article (7 references). The studies reported by Yasin et al. [3] and Zhou et al. [14] both reported that software engineering SRs did include grey literature. While none of these studies used exactly the same definition of grey literature, all confirmed the use of technical reports and PhD theses.

It is a fair criticism of the various guidelines produced by Kitchenham et al. (i.e., [15], [5] and [6], Section C) that they do not emphasise and explain the need to search grey literature clearly enough. However, looking in detail at Garousi et al.'s discussion of the value of grey literature in SE research [1] and, in particular, at their example based on the contribution of grey literature to a multivocal review [16], the issue is not whether grey literature is appropriate for inclusion in a systematic review, but whether information extracted from internet articles that do not conform with the Prague definition of grey literature should be included.

No guidelines for systematic reviews have ever proposed the use of social media posts or private communications as sources of primary studies for SRs, not only because social media posts were not a major information source when the guidelines were developed, but also because SRs are intended to aggregate empirical evidence from primary studies which have been subjected to critical assessment and which are expected to remain available in the public domain in the long term.

Nonetheless, Garousi et al. [1] raise the question as to whether social media posts should be regarded as grey literature and should, therefore, be considered suitable as sources for systematic reviews. In particular, bearing in mind Garousi and Mäntylä's multivocal review identified information from blogs, we question whether blogs should be regarded as grey literature. The main problem with blogs is that they may not report primary studies. Primary studies need to be full reports of research projects including research questions, description of the empirical and analysis methods used, their results, and their limitations. This level of detail is necessary in order for any evidence they report to be properly assessed for rigour and validity.

Equally importantly, blogs are not usually "collected and preserved" as mentioned in the Prague definition (see Section 5). SRs and systematic mapping studies both have requirements for auditability, traceability, and reproducibility that can only be met by white literature and grey literature conforming to the Prague definition. In particular, readers of a systematic review or systematic mapping study should be able to:

1) access all the primary studies identified in the review;
2) link individual primary studies to each reported finding.

In their multivocal review, Garousi and Mäntlyä found only six sources that reported empirical evidence, and all of those sources were classified as being formal literature. The emphasis on research-based evidence in SR guidelines is necessary because different experiences have shown that personal opinions, even those expressed by recognised experts, can be wrong or outdated. For example:

- Linus Pauling (a double Nobel Laureate) was incorrect in concluding that vitamin C prevented the common cold. He missed five important studies that had non-significant results [17] (see page 6).
- The logo of the Cochrane Organisation is based on a meta-analysis of studies of the use of corticosteroids by pregnant women expected to deliver premature babies [18]. It is based on the first eight papers studying the issue that were published before 1984. However, it was not until 1993 that the Royal College of Obstetricians and Gynecologists advised its members to use corticosteroids in all appropriate cases [19].

Delay in adopting the use of corticosteroids resulted in many unnecessary infant deaths.

- At one time it was "obvious" to pediatricians that premature babies should be put into an oxygen-rich environment. An unfortunate side-effect of this treatment was that many such babies (including Stevie Wonder) went blind.
- In 2001, Boehm and Basili claimed that "Perspective-based reviews catch 35 percent more defects than nondirected reviews" [20]. In 2009, Ciolkowski [21] reported the results of a meta-analysis of perspective-based (PBR) experiments saying "Our main findings regarding team effectiveness of PBR include that there is not a clear advantage of PBR over other reading techniques".
- For many years, it was an article of faith among SE researchers and tool vendors that models were better than humans at estimating the effort required for software development. Jørgensen's systematic review pointed out that aggregated empirical evidence did not support this view (see [22] and the subsequent update [23]). About a third of the empirical studies suggested model estimates were better than those made by humans, a third suggested not much difference in estimate accuracy, and the remaining studies suggested estimates from humans outperformed those from models.

It is also important that researchers and practitioners be aware of any limitations concerning the evidence supporting their current practices. Tatsioni et al. [24] report that "Claims from highly cited observational studies persist and continue to be supported in the medical literature despite strong contradictory evidence from randomized trials". This means that if current practice is based on studies of low rigour, such as expert opinion, it can be difficult to get practitioners to reconsider their views and accept contradictory findings from more rigorous studies. This also occurs in the context of software engineering. Devanbu et al. [25] report that practitioner beliefs are primarily based on personal experience, which can vary from project to project, but do not necessarily correspond to actual project evidence.

Although we do not believe that social media posts (or personal communications) report primary studies, we agree with Garousi and his colleagues that they can provide important information about new software engineering ideas and methods. Social media materials from OSS projects and Q&A posts have been widely used as source material for software engineering primary studies. In this paper, we discuss our view of how information from blogs and personal communications can be used both to examine the relevance of SR findings and also to support evidence-based software engineering in Section 7.

4 Representing the Practitioner Viewpoint

It is critical for systematic reviews that aim to provide advice to practitioners to include information from field studies. Curtis et al. [26] describe large-scale software development as a complex system involving individual programmers, the teams in which they work, the projects on which they work, the organisation that employs them, and the business sector in which the organisation does business. Laboratory experiments and small-scale validation studies that remove software engineering activities from their natural environment, cannot provide accurate assessments of the likely impact of a new technique when it is introduced into an industrial software production environment. We would argue that systematic reviews that do not include industry field studies can only provide weak evidence regarding the benefit of a new technique. Thus, another fair criticism of current SE guidelines for systematic reviews is that they do not make the importance of field studies clear enough, even though the medical guidelines on which the SE systematic review guidelines were initially based, considered only field studies of interventions to be admissible evidence. Results from animal experiments or laboratory experiments would *not* be considered for inclusion as candidate primary studies.

Given their view that blogs are grey literature that can be incorporated as primary studies in systematic reviews, Garousi and Mäntylä [16], argue that information from blogs provides an appropriate means of incorporating the viewpoints of practitioners into systematic reviews. However, blogs are not the *only* means capable of addressing the practitioner's viewpoint. Usually, any good quality industrial field study or case study should be able to help ensure that the findings of a systematic review will reflect practitioner values and priorities. For example, Budgen et al. [27], [28] selected 49 SE systematic reviews (from a set of 276) that included findings relevant to teaching about SE practice. They analysed 48 data sets used by these. In many cases, the primary studies were either explicitly or implicitly conducted in industry settings. Although the origin and form of the data from the primary studies were not always clearly reported, they were confident that 23 of the secondary studies were based mainly on industry studies and that a further 18 almost certainly included industry studies. Overall, it seems that the problem is not that SE systematic reviews exclude industry studies, it is more that SE researchers do not perform enough field studies and do not report the findings of such studies clearly enough. Nor do systematic reviewers always give enough emphasis to field studies in their analyses.

From the viewpoint of SRs, problems with incorporating the practitioner perspective arise from the fact that analysis of such information often leads to qualitative findings which cannot directly be aggregated with findings from quantitative research. However, there are methods for aggregating qualitative findings from different primary studies (for an overview of such approaches, see [29] and [30]). In addition, the Cochrane Handbook recommends using mixed-method reviews to include results from both quantitative and qualitative studies [31]. A mixed-method review is based on using aggregated qualitative findings to interpret and explain the results obtained from aggregated quantitative findings. Thus, the results of both qualitative aggregation and quantitative aggregation are kept separate (and can, therefore, be upgraded independently), but the findings from each aggregation are *compared* to provide more nuanced overall findings and recommendations. This is discussed further in Section 6.

5 Problems with Analysing Blogs

All research has inherent limitations, and furthermore, individual empirical research projects may not be of high quality. These problems, together with methods of addressing them, are regularly discussed in the software engineering literature (see, for example, [32], [33], [34], [35], [36], [37]). Systematic review guidelines from all disciplines acknowledge these issues, which is why they place strong emphasis on evaluating the methodological rigour of an SR and the risk of bias that may arise from individual primary studies.

Analysing blogs also suffers from limitations. Two important challenges are:

- *Bias*. This arises for two related reasons i) blog authors may have unstated vested interests and ii) they do not always represent the viewpoint of software engineering practitioners, because they may be produced by managers, consultants or tool vendors. For example, the list of bloggers reported by Rainer and Williams [38] includes influential and experienced software experts, but these are not *typical* software engineering practitioners. Furthermore, it is not clear that the existence of such biases will be recognised by readers, in particular students, who too readily assume that most internet material is trustworthy [39].
- *Lack of provenance*. Social media posts and private communications do not usually observe the need to cite original sources nor to respect copyright relating to graphics. For example, Garousi and Mäntylä [16] report that they found a similar graph identifying the return on investment from test automation in two different sources. This means that the notion of identifying independent pieces of evidence cannot be guaranteed, and using frequency counts to identify the importance of specific issues becomes misleading/valueless. Also, unlike the case for SRs based on primary studies [40], there is no accepted procedure for updating SRs that integrate social media posts and private communications with archival empirical studies.

For any research reports submitted to scholarly journals, there is a reasonable expectation that researchers have adhered to basic scientific principles, such as avoiding plagiarism, adhering to good practice in the conduct of their research and reporting any external research funding. Furthermore any researchers found to have violated these principles will be censured. No such expectations apply to social media posts. In addition, SRs based on primary studies are able to detect researcher bias. For example, Shepperd et al. [41] found that the outcome of defect prediction models was much more strongly related to the research group than the different prediction methods. Also, in his meta-analysis of perspective-based reading (that included informally published primary studies), Ciolkowski [21] reported that "Studies where the principle investigator had been involved in the initial PBR study (i.e., Basili 1996) or that use the original material set, tend to produce positive results, while the rest of the studies tend to produce negative results".

In general, assessing whether a blog is trustworthy is much more difficult than for a conventional research report because they seldom provide sufficient information to properly assess the risk of bias associated with their claims[3]. Nonetheless, Garousi et al.'s checklist in Table 7 [9] is a good contribution to the discussion of quality assessment of blogs.

In addition to the above issues, SE researchers have pointed out that there is an issue concerning the transience of online material [2], because there is no guarantee that the cited blogs or private communications will remain accessible in the long-term. For example, Garousi and Mäntylä [16] cited 46 internet articles and white papers using URL addresses, but, as of 25th May 2021, Kitchenham and Madeyski independently confirmed that only 19 were still accessible. A *partial* solution to this problem, is to use the Wayback Machine, available at https://archive.org. Using the Wayback Machine, it is enough to enter the failing URL, and if the archive includes that URL, which (unfortunately) is not always the case, it is possible to reach the missing reference. In our own example, we were able to retrieve 14 of the 27 references to grey literature that were previously unavailable. In addition, in one case, we were able to determine that a blog had been moved and to obtain its current location, so we were able to find a total of 15 missing references (56%). If researchers use the Wayback Machine as an integral part of the documentation of their data extraction and analysis process, it can completely overcome the problem of transience of the source material. However, such a solution may also lead to conflicts with copyright laws, while still not guaranteeing the long-term accessibility of the information to third parties, in particular, *readers* of the SR. Thus, the possible transience of blogs means that unlike grey literature conforming with the Prague definition, they represent a threat to the reproducibility of aggregated data. For example, it is clear from our example, that the results reported by Garousi and Mäntylä [16] are no longer fully auditable, traceable or reproducible by third party readers.

6 Guidelines for Grey Literature and Multivocal Reviews

Garousi et al. propose new guidelines for grey literature and multivocal reviews [9]. They acknowledge that current guidelines for conducting systematic reviews recommend including grey literature, however, their Figure 3 suggests that the use of grey literature is not addressed by current SR guidelines.

This contradiction has arisen because the reference to *grey literature* in SE systematic review guidelines is based on the definition of *grey literature* in the Luxembourg definition. Although the Luxembourg definition has been updated, the current Prague definition of grey literature is still appropriate for SRs. If the label *grey literature* is extended to forms of social media material that do not conform to the Prague definition, then we agree that SR guidelines do not address such material. However, we argue that SRs are not intended to incorporate such material. SRs and mapping studies require sources that are permanently available to third parties

3. Many well-respected software engineering experts in industry and academia publish blogs and we would expect to find valuable advice and ideas in their blogs. However, we would look to their books and more formally published articles for full reports of primary studies.

(i.e., individuals that are neither the authors of the material nor the SR authors).

With respect to multivocal studies, in the case of SRs (as opposed to mapping studies), the primary studies should report *empirical evidence* about the topic of interest. However, in their multivocal review [9], Garousi and Mäntylä found no sources reporting evaluation studies outside of the formal literature, so there is no evidence that social media material will provide additional value in the context of evaluation-based SRs.

However, in various studies, Garousi et al. have demonstrated that blog posts can contain opinions, ideas, and experiences of value to both practitioners and academics [42], [43], [16], [44]. Thus, there is a clear need to provide guidelines for searching and analysing social media posts. However, there is a good reason why a study aimed at aggregating blogs should not usually be regarded as being a form of secondary study. A blog should only be included as a primary study in a systematic review if it describes a well-conducted empirical study *which is not formally published elsewhere* and which is likely to be available in the long term.

In our opinion, reviewing a blog authored by a specific individual and extracting comments related to our own research questions is similar to analysing an unstructured interview. In contrast to systematic reviews and systematic mapping studies, readers of studies that have been based on unstructured interviews do not expect that the individual interview scripts will be available for them to read, nor do they expect to be able to confirm the relationship between the scripts and individual findings.

Qualitative primary studies have a requirement for methodological repeatability (being able to repeat the study methodology with different contexts, participants, or sources), but not reproducibility (being able to trace findings from the original study to each individual source). Thus, we agree with J. Adams et al. [13] who suggest that analysing social media information is akin to conducting a primary study. From this viewpoint, the methodology required to aggregate information from social media sources such as blogs should be based on qualitative research methodologies, not secondary study methodologies such as the systematic review methodology or systematic mapping study processes.

It would be less confusing and more accurate if we were to avoid the term *grey literature review* and refer to studies that survey information from social media sources by referring directly to the type of social media, and the nature of the study. For example, if a study examines blogs to identify opinions about the benefits and risks of the DevOps approach, it should use a title such as "Risk and Benefits of DevOps: A Survey of Blogs". This is exactly the way researchers studying Q&A sites identify the specific site, such as StackOverflow, in the title of their studies, see for example, [45], [46], [47].

The main difference between standard qualitative study methodology and blog surveys is that qualitative studies tend to emphasise the information elicitation and analysis processes more than the identification of participants, while blog surveys concentrate more on the identification of appropriate blogs. The emphasis on searching for appropriate information seems to be the rationale for trying to adopt the SR and systematic mapping methodology. However, the processes that can be used to extract and analyse data from textual material or videos actually require qualitative analysis methods.

A blog survey could be eligible for inclusion as a *primary study* in a *qualitative* systematic review. However, as shown by Garousi et al. [9] in their Table 7, it would require primary study quality evaluation criteria different from those used in a more traditional qualitative study. In addition, it should be noted that the original SR guidelines for SE (i.e., [15] and [5]) did not discuss the issue of aggregating qualitative primary studies. The most recent guidelines, reported in [6], have attempted to rectify this omission.

In Section 2, we mentioned that company and industry private communications are sometimes included in extended definitions of grey literature. However, we would not expect information of this sort to form the main source material for any empirical study. It is generally used in field studies as a source of triangulation data to validate other data sources and/or provide contextual information to help explain other qualitative or quantitative study findings.

With respect to the guidelines for multivocal reviews, the ambiguities we have identified with respect to the term *grey literature review* suggest that the focus of the guidelines in [9] needs to be revised to refer explicitly to blog surveys as qualitative primary studies. The fact that the model in Figure 7, as presented in [9], is adapted from a model of the systematic mapping process does not imply that the activity being modelled is a form of systematic review. It merely reflects the fact that at a high level of abstraction, all empirical research projects have broadly similar processes based on defining goals and research questions, identifying appropriate research methods, obtaining data to address those research questions, and analysing the data.

In particular, Guideline 13 from [9], which concerns data synthesis, needs to be refined. We have argued that primary studies found in grey literature (using the Luxembourg or Prague definition) can be included in systematic reviews in the same way as a primary study. However, we agree with R. Adams et al. [12] that surveys of blogs must be treated as sources of information with low credibility, which should *not* be formally aggregated with other sources of evidence. Nonetheless, there is undoubtedly value to be had from the results of blog surveys, and the main issue is how to put such results to use in order to assist software practitioners. We discuss this in Section 7.

Overall, the use of surveys of blogs raises non-trivial methodological issues. The guidelines in [9] provide the basis for developing useful methodological guidelines for undertaking such a study (in particular, their Table 7 identifies information similar to the demographic data and contextual information used in qualitative surveys). However, all of the guidelines need to be reviewed and assessed against qualitative research guidelines.

There also needs to be consideration of the ethical issues associated with the use of social media such as blogs and vlogs. For example, it may be difficult to distinguish malicious and untrue comments from fair and reasonable comments. Thus, there is a danger that an academic publication including unvetted blog content could add legitimacy to untrue or malicious comments. In addition, use of the Wayback Machine can also be ethically questionable given

the current debate about issues such as the right to be forgotten[4].

7 Using the Findings of Blog Surveys

In this section, we discuss how the findings from blog surveys can be used:

1) as input to *mixed methods* reviews;
2) as input to the fourth stage of evidence-based software engineering (EBSE) [6], [10].

We also consider how findings from all the information sources identified in Figure 1 can be incorporated into the EBSE framework to support industry-relevant guidelines and recommendations.

7.1 Mixed-Methods Reviews

Mixed-methods reviews are recommended by the Cochrane Handbook [31] as a means of comparing findings from quantitative systematic reviews with findings from qualitative studies and qualitative systematic reviews. In the context of blog surveys, we need to be able to:

- Report the credibility of findings from an SR based on primary studies separately from the findings from blog surveys. It is critical that readers know the provenance of all recommendations, so they can properly judge their credibility.
- Compare the findings of systematic reviews with findings from blog surveys in order to look for agreements, disagreements, and content missing from the different sources.

Comparing SR results with data from blog surveys potentially presents us with a powerful method to assess the practical relevance of SE research:

1) If we have agreement between findings from a systematic review and findings from blogs, then we can have some confidence that our findings can be trusted.
2) If the findings are inconsistent, we should give preference to the SR results, but investigate possible contextual factors that might explain the inconsistencies.
3) If there are blog findings but no corresponding SR findings, we have a potentially important topic that would benefit from more formal study and evaluation.
4) If SR findings relate to topics that are not mentioned in any blogs, the SR may be reporting an issue that is of little relevance to industry.

Figure 5 of Garousi et al.'s book chapter [48] shows how our view of the use of blogs in SRs differ, but can easily be reconciled. In their original paper, Garousi and Mäntylä [9] present a graphic that identifies the number of times a specific topic was mentioned in the sources they analyzed. In contrast, Figure 5 in Garousi et al. [48] shows the number of primary studies that discuss a specific topic and compares it with the number of blogs and industry white papers that mention the same topic. This is a much better way of representing SR and social media survey findings, although we would include any white papers that report primary studies and conform with the Prague definition in the formal literature. If information from the blogs and other white papers were treated as findings from a single primary study, as we suggest, each finding from the survey would add only a single count to each of the topics mentioned (perhaps with the details reporting the percentage of blogs that mentioned the topic). If test oracles and the development process were not mentioned by any white papers allocated to the formal literature, the diagram would still identify these as issues that do not appear to be directly addressed by reports in the formal literature concerning when and what to automate[5]. In addition, if the findings are linked to the individual primary studies and the other sources, it would be clear which findings are intended to be reproducible and which are not.

7.2 Using Blog Survey Results as Input to EBSE

Kitchenham et al. [6], [10] reworded the evidence-based medicine process to reflect the software engineering context. The first three steps involve i) converting the need for information into an answerable question, ii) tracking down the best evidence to answer that question, and iii) critically appraising that evidence. Systematic reviews are one method of addressing these three steps. The fourth step concerns integrating the critical appraisal with software engineering expertise and stakeholders' values. Findings from blog surveys are one source of software engineering expertise that can be considered during this stage[6].

7.3 Using Information Sources and Study Types in the EBSE Framework

Figure 2 summarises our view of how findings from different information sources and different study types contribute to Evidence-Based Software Engineering (EBSE) [6], [10]. Based on the discussion in this paper, Figure 2 shows the scope of each type of study, highlights the different types of information used in each study type, and identifies the different ways the information sources are analysed. Finally, it identifies EBSE as a means of bringing different findings together in order to produce industry-relevant recommendations for SE practice and decision making.

8 Conclusions and Recommendations

Current SR guidelines have been criticised for failing to include grey literature and the viewpoint of practitioners. We agree that current SR guidelines do not emphasise the critical importance of including field studies in those systematic reviews aiming to influence industry practice. However, they do not exclude the use of grey literature that conforms to the Prague definition. An important point about the Prague definition is that it emphasizes literature from academia and industry that has been *collected and preserved*.

4. See https://en.wikipedia.org/wiki/Right_to_be_forgotten

5. It is important to note that lack of findings from the formal literature related to specific questions does not imply that those issues are ignored in software engineering research. Garousi and Mäntylä [16] identify a systematic review addressing the oracle problem, and several others that discuss testing in the context of specific development processes.

6. The introduction of context and personal opinion during this stage of EBSE is one justification for using the term *evidence-informed* rather than *evidence-based*, as is becoming the norm in other disciplines.

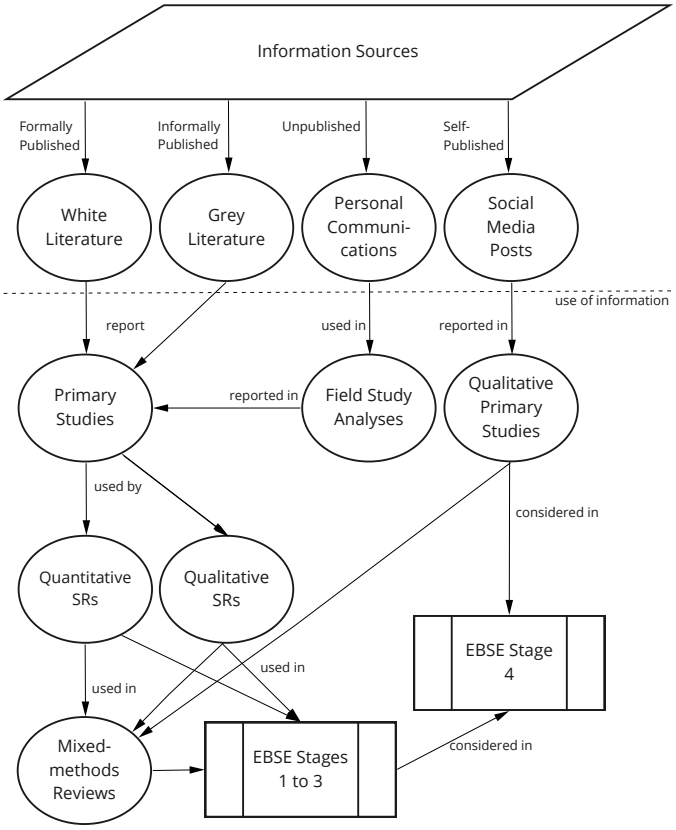

Fig. 2. How white and grey literature, personal communications and social media posts can be used in Evidence-Based Software Engineering

The emphasis placed in the Prague definition on collection and preservation of grey literature is important for systematic reviews and mapping studies because they have a requirement to support auditability, traceability and reproducibility. In the event that a transient social media post fulfils the eligibility criteria to be included in a systematic review as a primary study, it is the responsibility of SR author(s) to ensure the long term availability of the material to third parties. Ignoring the requirement for auditability, traceability and reproducibility would cause SRs and mapping studies produced in SE to be substantially weaker than those in other disciplines. There might also be problems adapting and using potentially useful tools that have been developed in other disciplines.

Thus, the answer to our high level research question is that the current SR guidelines do not need any major revisions for grey literature and multivocal reviews. Candidate primary studies found in the grey literature that conform with the Prague definition should be treated in exactly the same way as any other primary studies and can include industry-based field studies as well as academic experiments. However, the guidelines should be read with the following issues in mind:

- The term *grey literature* refers to the Prague definition of grey literature which emphasizes literature that has been collected and preserved in order to support SR reproducibility.
- Searching for appropriate grey literature is more difficult than searching for conventionally published articles and sources. Authors need to consider citation searching of identified primary studies (i.e., snowballing), direct approaches to subject experts, searching sources that catalogue PhD and MSc theses, searching sources such as archive sites and protocol registration sites, as well as using Google Scholar.
- Field studies are essential for evaluating software engineering methods and tools.

Discussing different information types, we argue that:

- Using the term *grey literature* to include concepts that differ in essence, not just in degree, can lead to a misunderstanding of how information from different types of literature can be used.
- Studies of internet material such as Q&A fora and OSS project information should be treated as primary studies and, in most cases, we would suggest treating surveys of blogs and vlogs in the same way. Thus, guidelines for aggregating these should be based on qualitative research guidelines rather than secondary study guidelines. However, as indicated in Figure 1, they can still be very useful for both academics and practitioners.
- Government and industry white papers and academic technical reports would usually be treated as grey literature conforming to the Prague definition and treated in the same way as formal literature.
- Personal communications from academia, industry, and government sources provide ancillary information to allow data triangulation and to provide contextual information about other findings. They are unlikely to be the sole basis for any primary or secondary study.

To address the question as to how different information sources can be used in software engineering research, we present our categorisation of different information sources in Figure 1 and their use in subsequent academic research in Figure 2.

With respect to input drawn from social media posts of all types, we do not dispute its potential value, particularly for identifying new ideas and suggestions that could be the inspiration for rigorous empirical studies. In particular, findings from surveys of blogs can provide valuable information about issues of concern to industry, while information from Q&A sites, as well as providing direct answers to practitioners' coding and design problems, should be of relevance to the authors of design and coding handbooks and training materials, as well as to software engineers involved in developing code and considering the use of design patterns (see for example [47]).

We conclude with the following recommendations for SE researchers, based on the arguments presented in this paper.

Recommendation 1: Clearly distinguish information obtained from grey literature conforming with the Prague definition from information obtained from other social media material.

Recommendation 2: Do not arbitrarily exclude primary studies obtained from grey literature studies from inclusion in SRs.

Recommendation 3: Only systematic reviews that include rigorous field studies or large-scale (realistic) empirical evaluations should make recommendations regarding industry SE practice.

Recommendation 4: Use the term *survey*, not *grey literature review*, to refer to any study aimed at aggregating personal opinions derived from blogs.

Recommendation 5: Use information from studies that aggregate blogs to support the interpretation of systematic reviews, and/or the fourth step in the EBSE process [6], [10].

Recommendation 6: Use information from private communication channels to support validation of qualitative data and interpretation of quantitative study findings.

Recommendation 7: Ensure that any social media material reporting a primary study will be permanently and legally available to the SR readers.

Acknowledgment

We thank Sebastián Pizard, of the School of Engineering, Universidad de la República, Montevideo, Uruguay, for pointing out the ethical issues associated with the use of social media posts in academic publications. We also thank the anonymous reviewers for helping us to refine our arguments and improve the presentation of our ideas.

References

[1] V. Garousi, M. Borg, and M. Oivo, "Practical relevance of software engineering research: synthesizing the community's voice," *Empirical Software Engineering*, vol. 25, pp. 1687—1754, 2020.
[2] F. Kamei, I. Wiese, C. Lima, I. Polato, V. Nepomuceno, W. Ferreira, M. Ribeiro, C. Pena, B. Cartaxo, G. Pinto, and S. Soares, "Grey literature in software engineering: A critical review," *Information and Software Technology*, vol. 138, p. 106609, 2021.
[3] A. Yasin, R. Fatima, L. Wen, W. Afzal, M. Azhar, and R. Torkar, "On using grey literature and google scholar in systematic literature reviews in software engineering," *IEEE Access*, vol. 8, pp. 36 226–36 243, 2020.
[4] H. Zhang, X. Zhou, X. Huang, H. Huang, and M. A. Babar, "An evidence-based inquiry into the use of grey literature in software engineering," in *42nd International Conference on Software Engineering (ICSE)*, 2020, pp. 1422–1434.
[5] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele University, UK, Tech. Rep. EBSE-2007-01, 2007.
[6] B. Kitchenham, D. Budgen, and P. Brereton, *Evidence-Based Software Engineering and Systematic Reviews*. CRC Press, 2016.
[7] R. K. Yin, *Case Study Research Design and Methods*, 5th ed. Sage Publications Inc., 2014.
[8] J. Schöpfel, "Towards a Prague Definition of Grey Literature," in *Twelfth International Conference on Grey Literature: Transparency in Grey Literature, Grey Tech Approaches to High Tech Issues, 6-7 December 2010*, 2010.
[9] V. Garousi, M. Felderer, and M. V. Mäntylä, "Guidelines for including grey literature and conducting multivocal literature reviews in software engineering," *Information and Software Technology*, vol. 106, pp. 101–121, 2019.
[10] B. Kitchenham, T. Dybå, and M. Jørgensen, "Evidence-based software engineering," in *Proceedings of ICSE 2004*. IEEE Computer Society Press, May 2004, pp. 273–281.
[11] T. L. Rucinski, "The elephant in the room: Toward a definition of grey legal literature," *Law Library Journal*, vol. 107, no. 4, pp. 543–559, 2015.
[12] R. J. Adams, P. Smart, and A. S. Huff, "Shades of grey: Guidelines for working with the grey literature in systematic reviews for management and organizational studies," *International Journal of Management Reviews : IJMR*, vol. 19, no. 4, pp. 432–454, 2017.
[13] J. Adams, F. C. Hillier-Brown, H. J. Moore, A. A. Lake, V. Araujo-Soares, M. White, and C. Summerbell, "Searching and synthesising 'grey literature' and 'grey information' in public health: critical reflections on three case studies," *Systematic Reviews*, vol. 5, no. 1, pp. 164–164, 2016.
[14] X. Zhou, Y. Jin, H. Zhang, S. Li, and X. Huang, "A map of threats to validity of systematic literature reviews in software engineering," in *2016 23rd Asia-Pacific Software Engineering Conference (APSEC)*. IEEE, 2016, pp. 153–160.
[15] B. Kitchenham, "Procedures for undertaking systematic reviews," Joint Technical Report Keele and Durham Universities, Tech. Rep., 2004.
[16] V. Garousi and M. V. Mäntylä, "When and what to automate in software testing? A multi-vocal literature review," *Information and Software Technology*, vol. 76, pp. 92—117, 2016.
[17] M. Petticrew and H. Roberts, *Systematic Reviews in the Social Sciences: A Practical Guide*. Wiley-Blackwell, 2006.
[18] J. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, and V. Welch, Eds., *Cochrane Handbook for Systematic Reviews of Interventions Version 6.2*. Wiley, 2021, available from www.training.cochrane.org/handbook.
[19] P. A. Crowley, "Antenatal corticosteroid therapy: A meta-analysis of the randomized trials, 1972 to 1994," *American Journal of Obstetrics and Gynecologyrican Journal of*, vol. 173, no. 1, pp. 322—335, 1995.
[20] B. Boehm and V. R. Basili, "Software defect reduction top 10 list," *IEEE Computer*, pp. 135–137, 2001.
[21] M. Ciolkowski, "What do we know about perspective-based reading? an approach for quantitative aggregation in software engineering," in *Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement*, ser. ESEM '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 133–144. [Online]. Available: http://dx.doi.org/10.1109/ESEM.2009.5316026
[22] M. Jørgensen, "A review of studies on expert estimation of software development effort," *Journal of Systems and Software*, vol. 70, no. 1-2, pp. 37–60, 2004.
[23] ——, "Forecasting of software development work effort: Evidence on expert judgement and formal models," *Int. Journal of Forecasting*, vol. 23, no. 3, pp. 449–462, 2007.
[24] A. Tatsioni, N. G. Bonitsis, and J. P. A. Ioannidis, "Persistence of contradicted claims in the literature," *JAMA*, vol. 298, no. 21, pp. 2517–2526, 2007.
[25] P. Devanbu, T. Zimmermann, and C. Bird, "Belief & Evidence in Empirical Software Engineering," in *38th International Conference on Software Engineering (ICSE)*, 2016, pp. 108–119.
[26] B. Curtis, H. Krasner, and N. Iscoe, "A Field Study of the Software Design Process for Large Systems," *Communications of the ACM*, vol. 31, no. 11, pp. 1268–1287, 1988.
[27] D. Budgen, P. Brereton, N. Williams, and S. Drummond, "The contribution that empirical studies performed in industry make to the findings of systematic reviews: A tertiary study," *Information and Software Technology*, vol. 94, pp. 234 – 244, 2018.
[28] D. Budgen, P. Brereton, N. Williams, and S. Drummond, "What support do systematic reviews provide for evidence-informed teaching about software engineering practice?" *e-Informatica Software Engineering Journal*, vol. 14, pp. 7—60, 2020.
[29] J. L. Harris, A. Booth, M. Cargo, K. Hannes, A. Harden, K. Flemming, R. Garside, T. Pantoja, J. Thomas, and J. Noyes, "Cochrane Qualitative and Implementation Methods Group guidance series paper 2: methods for question formulation, searching, and protocol

development for qualitative evidence synthesis," *Journal of Clinical Epidemiology*, vol. 97, pp. 38—48, 2018.
[30] J. Noyes, A. Booth, K. Flemming, R. Garside, A. Harden, S. Lewin, T. Pantoja, K. Hannes, M. Cargo, and J. Thomas, "Cochrane Qualitative and Implementation Methods Group guidance series paper 3: methods for assessing methodological limitations, data extraction and synthesis, and confidence in synthesized qualitative findings," *Journal of Clinical Epidemiology*, vol. 97, pp. 69—58, 2018.
[31] J. Noyes, A. Booth, M. Cargo, F. K, A. Harden, J. Harris, R. Garside, K. Hannes, T. Pantoja, and J. Thomas, *Cochrane Handbook for Systematic Reviews of Interventions Chapter 21: Qualitative evidence*, J. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. P. MJ, and V. Welch, Eds. Cochrane, 2021, vol. version 6.2, www.training.cochrane.org/handbook.
[32] B. Kitchenham, S. Pfleeger, L. Pickard, P. Jones, D. Hoaglin, K. El Emam, and J. Rosenberg, "Preliminary guidelines for empirical research in software engineering," *IEEE Transactions on Software Engineering*, vol. 28, no. 8, pp. 721–734, 2002.
[33] M. Shepperd, D. Bowes, and T. Hall, "Researcher bias: The use of machine learning in software defect prediction," *IEEE Transactions on Software Engineering*, vol. 40, no. 6, pp. 603–616, 2014.
[34] M. Jørgensen and E. Papatheocharous, "Believing is seeing: Confirmation bias studies in software engineering," in *41st Euromicro Conference on Software Engineering and Advanced Applications*, 2015.
[35] M. Jørgensen, T. Dybå, K. Liestøl, and D. I.K.Sjøberg, "Incorrect results in software engineering experiments: How to improve research practices," *The Journal of Systems and Software*, vol. 116, pp. 133–145, 2016.
[36] M. Shepperd, N. Ajienka, and S. Counsell, "The role and value of replication in empirical software engineering results," *Information and Software Technology*, vol. 99, pp. 120–132, 2018.
[37] B. Kitchenham, L. Madeyski, and P. Brereton, "Problems with Statistical Practice in Human-Centric Software Engineering Experiments," in *Proceedings of the Evaluation and Assessment on Software Engineering*, ser. EASE '19. New York, NY, USA: ACM, 2019, pp. 134–143.
[38] A. Rainer and A. Williams, "Using blog articles in software engineering research: benefits, challenges and case–survey method," in *25th Australasian Software Engineering Conference (ASWEC)*. IEEE Computer Society, 2018, pp. 201—209.
[39] L. Graham and P. T. Metaxas, ""Of Course It's True; I Saw It on the Internet!": Critical Thinking in the Internet Era," *Commun. ACM*, vol. 46, no. 5, p. 70–75, May 2003. [Online]. Available: https://doi.org/10.1145/769800.769804
[40] P. Garner, S. Hopewell, J. Chandler, H. MacLehose, E. A. Akl, J. Beyene, S. Chang, R. Churchill, K. Dearness, G. Guyatt, C. Lefebvre, B. Liles, R. Marshall, L. Martínez García, C. Mavergames, M. Nasser, A. Qaseem, M. Sampson, K. Soares-Weiser, Y. Takwoingi, L. Thabane, M. Trivella, P. Tugwell, E. Welsh, E. C. Wilson, and H. J. Schünemann, "When and how to update systematic reviews: consensus and checklist," *BMJ*, vol. 354, 2016. [Online]. Available: https://www.bmj.com/content/354/bmj.i3507
[41] M. Shepperd, D. Bowes, and T. Hall, "Researcher bias: The use of machine learning in software defect prediction," *IEEE Transactions on Software Engineering*, vol. 40, no. 6, pp. 603–616, 2014.
[42] V. Garousi, M. Felderer, and M. V. Mäntylä, "The need for multivocal literature reviews in software engineering: complementing systematic literature reviews with grey literature," in *Proceedings of EASE'16*. ACM, 2016.
[43] V. Garousi, M. Felderer, and T. Hacaloğlu, "Software test maturity assessment and test process improvement: A multivocal literature review," *Information and Software Technology*, vol. 85, pp. 16—42, 2017.
[44] V. Garousi and A. Rainer, "Grey literature versus academic literature in software engineering: A call for epistemological analysis," *IEEE Software*, 2020.
[45] H. Zhang, S. Wang, T.-H. Chen, and A. E. Hassan, "Reading Answers on Stack Overflow: Not Enough!" *IEEE Transactions on Software Engineering*, vol. 47, no. 11, pp. 2520–2533, 2021.
[47] X. Liu and H. Zhong, "Mining StackOverflow for Program Repair," in *2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 2018, pp. 118–129.
[46] D. Van Der Linden, E. Williams, J. Hallett, and A. Rashid, "The Impact of Surface Features on Choice of (in)Secure Answers by Stackoverflow Readers," *IEEE Transactions on Software Engineering*, pp. 1–1, 2020.
[48] V. Garousi, M. Felderer, M. V. Mäntylä, and A. Rainer, "Benefiting from the grey literature in software engineering research," in *Contemporary Empirical Methods in Software Engineering*, M. Felderer and G. H. Travassos, Eds. Springer, 2020, ch. Benefiting from Grey Literature in Software Engineering Research.

Barbara Kitchenham is an Emeritus Professor in the School of Computing and Mathematics at Keele University in the UK. She has worked in software engineering for over 40 years both in industry and academia. She has published over 150 software engineering journal and conference papers. Her most recent research has focused on the application of evidence-based practice to software engineering. In 2019, she was awarded the IEEE Technical Committee Distinguished Women in Science & Engineering (WISE) Leadership Award.

Lech Madeyski is an Associate Professor & Deputy Head of the Department of Applied Informatics at Wroclaw University of Science and Technology, Poland. He has been a Visiting Researcher at Keele University (UK), Brunel University London (UK), and a Visiting Professor at Blekinge Institute of Technology (Sweden). His research focus is on empirical software engineering, data science in software engineering, reproducible research, robust statistical methods. He is a co-founder of *e-Informatica Software Engineering Journal* & International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE). He has published in prestigious journals including, e.g., *IEEE Transactions on Software Engineering*. He is an author of a book "Test-Driven Development: An Empirical Evaluation of Agile Practice" incl. meta-analysis of experiments.

David Budgen is an Emeritus Professor of Software Engineering in the Department of Computer Science at Durham University in the UK. His research interests include software design, design environments, health-care computing and evidence-based software engineering. He was awarded a BSc(Hons) in Physics and a PhD in Theoretical Physics from Durham University, following which he worked as a research scientist for the Admiralty and then held academic positions at Stirling University and Keele University before moving to Durham University in 2005. He is a member of the IEEE Computer Society, the ACM and the Institution of Engineering & Technology (IET), and is a Chartered Engineer (CEng).