Incorporating Altmetrics to Support Selection and Assessment of Publications During Literature Analyses

Yusra Shakeel
Otto-von-Guericke University
Magdeburg, Germany
shakeel@ovgu.de

Rand Alchokr
Otto-von-Guericke University
Magdeburg, Germany
rand.alchokr@ovgu.de

Jacob Krüger
Ruhr-University Bochum
Bochum, Germany
Otto-von-Guericke University
Magdeburg, Germany
jacob.krueger@rub.de

Gunter Saake
Otto-von-Guericke University
Magdeburg, Germany
saake@ovgu.de

Thomas Leich
Harz University of Applied Sciences
Wernigerode, Germany
METOP GmbH
Magdeburg, Germany
tleich@hs-harz.de

ABSTRACT

Background. The constantly increasing number of scientific publications poses challenges for researchers to monitor, select, and assess the publications relevant for their own research. Several guidelines for assessing publications manually during a literature analysis exist, with researchers proposing (semi-)automated techniques to facilitate such assessments. **Aims.** Still, research indicates that current techniques require further improvements to facilitate the analysis of large sets of publications. In this paper, we propose a semi-automatic technique with which we aim to improve in this direction by facilitating the selection and assessment of publications. **Method.** Our technique uses publicly available data of a publication, namely citation counts, article-level metrics, venue metrics, and altmetrics, to guide an analyst in assessing its relevance and impact. To evaluate the feasibility of our technique and the included metrics, we performed an experimental analysis to automatically assign ratings to the retrieved publications. **Results.** The results indicate that our technique can help an analyst in assessing publications, and reduce manual effort. Through our technique, we achieve an average accuracy of 53 % with a recall of 71 %. While precision (14 %) and F1-score (21 %) are—not surprisingly, due to the high number of irrelevant results returned by automatic searches in digital libraries—low, we see an improvement of these values for more recent reviews for which we could collect more complete data. However, some manual effort is still required for the final selection of papers. **Conclusions.** While it is not possible to achieve full automation for selecting and quality assessing publications, we can see that our metrics-based technique can be a helpful means to provide an initial rating for the analyst. Also, incorporating altmetrics seems to be a promising addition to rate comparably recent publications, helping researchers to further facilitate the execution of literature analyses.

EASE 2022, June 13–15, 2022, Gothenburg, Sweden

ACM ISBN 978-1-4503-9613-4/22/06.
https://doi.org/10.1145/3530019.3530038

CCS CONCEPTS

• **General and reference** → **Surveys and overviews**; • **Information systems** → *Document filtering*.

KEYWORDS

Literature analysis, Quality assessment, Altmetrics, PlumX

ACM Reference Format:
Yusra Shakeel, Rand Alchokr, Jacob Krüger, Gunter Saake, and Thomas Leich. 2022. Incorporating Altmetrics to Support Selection and Assessment of Publications During Literature Analyses. In *The International Conference on Evaluation and Assessment in Software Engineering 2022 (EASE 2022), June 13–15, 2022, Gothenburg, Sweden.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3530019.3530038

1 INTRODUCTION

In computer science, and particularly software engineering, the number of publications describing new techniques, tools, methods or empirical studies is constantly increasing. As a result, it takes researchers and practitioners more and more time to identify and assess those publications that are relevant for them; we refer to an analyst who conducts a literature analysis. Particularly time consuming is any systematic review, such as a systematic literature review or a systematic mapping study, in which an analyst aims to cover all publications related to a certain topic to provide an overview of the research that has been conducted [22, 49]. However, even if the analyst does not follow such systematic methods based on defined guidelines, they still have to perform similar steps, namely searching, selecting, and assessing publications.

The most reliable method for selecting a publication and assessing its quality remains reading it carefully. As guidance, researchers have proposed well-defined checklists [14, 22] to determine a publication's importance and quality, for example, considering reporting, rigor, credibility, and relevance. However, such a manual analysis requires considerable effort and time, especially when facing

a large—and steadily increasing—number of potentially relevant publications. To tackle this problem, researchers have proposed (semi-)automated techniques to determine a publication's relevance and quality, and thus reduce an analyst's workload [20, 29, 43].

Unfortunately, while the mapping study of Marshall and Brereton [29] shows that software-engineering researchers aim to automate particularly the selection and quality assessment, the success seems rather limited. Recently, Hassler et al. [20] and we [43] identified that existing techniques for systematic literature reviews need to be improved, particularly for these two steps—which are among the most desired feature of the research community. So, while promising immense benefits by reducing an analyst's workload and the time needed, existing techniques and tools seem to require further improvements with respect to guiding the selection and assessment of relevant publications.

In this paper, we propose a technique to assess the relevancy and impact of publications using their citations and bibliometrics; thus reducing the time required to perform a literature analysis. We remark that a fully automated tool that yields perfect results is impossible, as the tool would need to fully understand natural language and the analyst's intentions. So, we aim to support an analyst by providing guidance on what papers are more likely to be relevant for them. To achieve this goal, our technique utilizes citation relationships and combines them with bibliographic information (i.e., author contribution and venue metrics). Additionally, we integrate altmetrics, which are metrics that measure a publication's impact on various social media platforms [13, 17]. Consequently, we build on ideas of existing techniques [43], extend them, and integrate altmetrics as a new type of metrics. We argue that altmetrics reflect differently on the impact of a publication [40, 41]; assuming that especially well-crafted and high-impact publications are discussed and gain recognition faster than reflected in citations. In detail, our contributions are:

- We propose a technique to assess publications based on their citation links, meta data, and altmetrics.
- We evaluate our technique based on an empirical comparison against 10 existing systematic reviews, including systematic literature reviews and systematic mapping studies.
- We discuss how our technique and especially altmetrics reflect on a publication's importance.
- We provide an open-access dataset comprising our prototype and evaluation data.[1]

The results show that our technique achieves an accuracy of 53 % and a recall of 71 %, indicating that it can be particularly helpful to identify irrelevant publications. As we are working only with metrics and few of the initially found publications (even of high quality) are usually relevant for literature reviews, it is not surprising that precision and F1-score are rather low. Still, the results indicate that our technique and the considered metrics can guide analysts, and allow researchers to further facilitate literature analyses.

2 BACKGROUND AND GOALS

Literature analyses are an essential research method to consolidate the existing knowledge or evidence regarding a specific problem, allowing to critically analyze that knowledge and identify open gaps. However, the structure, goal, and presentation of literature analyses depend on the type of analysis and has changed over time. Various classifications of literature analyses have been proposed, including quantitative or qualitative systematic literature reviews, systematic mapping studies, tertiary studies, narrative reviews, development reviews, cumulative reviews, aggregative reviews, descriptive reviews, scoping reviews, meta-analyses, umbrella reviews, theoretical reviews, realist reviews, or critical reviews [21, 22, 35]. In this paper, we focus on *systematic literature reviews* and *systematic mapping studies* in software engineering, according to the guidelines of Kitchenham et al. [21]. While the goals of different types of analyses vary, and systematic literature reviews in software engineering are regularly criticized for reporting only publication statistics, we focus on these because they build on well-defined guidelines. More precisely, we are only concerned with the *systematic process* of identifying relevant publications. A systematic conduct is arguably a favorable property for any literature analysis to allow others to verify its quality, to replicate it, and to assess its completeness—essentially improving the trust in and usability of the results. So, a systematic, understandable conduct is important for any literature analysis [21, 24, 42, 49].

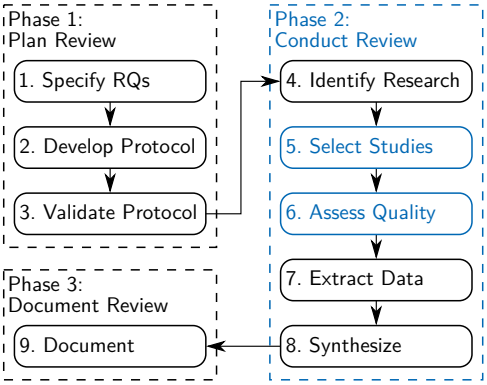

Figure 1: The systematic literature-review process adopted from Kitchenham et al. [21]. We highlight the activities we are concerned with in blue.

We display the phases and activities of a systematic literature review in Figure 1. Initially, the analyst must plan the review, defining what research questions they want to answer, developing a review protocol, and validating that protocol. This planning involves defining a search strategy, search terms, data sources, selection criteria, and quality criteria. In the second phase, the analyst conducts the review, essentially instantiating the defined protocol by identifying (i.e., searching), selecting, and quality assessing publications. Then, they extract the data that is relevant for answering their research questions and synthesize that data. Finally, the last phase is concerned with documenting all activities and results, potentially for publishing them.

Goals. In this paper, we are concerned with two activities within a literature analysis, namely; to select relevant studies and assess their quality (highlighted in blue in Figure 1). We remark that this refers to the quality assessment of an individual publication, as opposed to assessing the findings of the analysis itself [21]. So, we

[1] https://www.dropbox.com/s/f5cxmlbhfxf571p/Artifacts.zip?dl=0

aim to interpret the quality and impact a publication has on the scientific community as reflected, for instance, by citation counts and altmetrics. To efficiently select the most important publications is an essential activity for a systematic literature analysis, limiting the selection bias, improving the reliability of the results, and essentially enabling others to evaluate the value of the analysis. Due to the importance of this activity, detailed guidelines have been proposed, typically evaluating properties like study design, data collection, quality of findings, hypotheses, reporting, rigor, credibility, and relevance [14, 21, 22, 50]. Moreover, researchers have proposed semi-automated techniques to facilitate the selection of publications, and thus reduce the effort of conducting a literature analysis [16, 28, 36, 43]. Most of these techniques face limitations regarding the employed measurements, and are often focused on visual support—since a full automation of selecting and assessing publications is hardly reliable. This observation is supported by recent studies highlighting the importance of improving the tool support especially for selecting and assessing publications [20, 43].

With our technique, we intend to improve this situation, aiming to achieve the following goals:

- G_1 Improve an analyst's efficiency by indicating the most important publications for their analysis.
- G_2 Improve an analyst's understanding by providing meaningful metrics on a publication, prior to viewing the full text.

The outcome of our technique are scorings that indicate a publication's relevance and impact based on the defined search string and additional meta data. We build particularly on the proposal of Ponsard et al. [36] who show that citation links can also be helpful for this purpose. As a result, we decided to combine analyses of a publication's meta data (e.g., citations, authors, venue), with altmetrics, which we describe in the next paragraph. The data we use is publicly available, making our technique highly accessible.

Altmetrics. Introduced in 2010, altmetrics are an alternative to traditional publication metrics, such as, citation count and venue impact factor, for assessing the impact of publications [13, 17]. Altmetrics build on usage data of a publication, namely the number of downloads, views, saves, and how the audience engages with the publication on social-media platforms, such as, Facebook and Twitter. The most important feature of such metrics is the immediate feedback through interactions on the Web that can be accumulated in a short period of time, contrary to citations that require a longer period to accumulate [37]. Although the research community is not completely convinced regarding the accuracy of altmetrics to evaluate quality (e.g., they can be easily manipulated [11, 12]), there is still evidence on their usefulness in terms of speed, diversity, ease of access, and coverage of different platforms [2, 8, 31, 33]. So, we believe that altmetrics can *support* literature analyses to obtain an automated, initial assessment that accounts for missing scientific impact of publications [40, 41]. More precisely, if a publication gets a number of tweets, mentions or downloads, it has a higher probability of being cited in the future [15, 47]. Consequently, such interactions may indicate the importance of new research better than traditional metrics, and maybe better at all (e.g., researchers interacting with high-quality publications that are outside of their domain, and thus not cited by them). Due to the increasing popularity of altmetrics, several tools and APIs have been developed

```
@article{Steinmacher2015,
        author = "Steinmacher, Igor and Graciotto Silva, Marco A.
                and Gerosa, Marco A. and Redmiles, David F.",
        title = "A Systematic Literature Review on the Barriers
                Faced by Newcomers to Open Source Software
                Projects",
        journal = "Information and Software Technology",
        year = "2015",
        doi = "https://doi.org/10.1016/j.infsof.2014.11.001",
        volume = "59",
        pages = "67--85",
        keywords = "Barriers to entry, Joining, Newcomers,
                Onboarding, Open source software, Systematic
                literature review",
        abstract = "Context. Numerous open source software
                projects are based on volunteers collaboration
                and require a continuous influx of newcomers for
                their continuity. Newcomers face barriers that
                can lead them to give up. These barriers hinder
                both developers willing to make a single
                contribution and those willing to become a
                project member. Objective. This study aims to
                identify and classify the barriers that newcomers
                face when contributing to open source software
                projects. [...]",
        cited-by = "120",
        SJR = "0.606",
        SNIP = "2.389",
        CiteScore = "8.6",
        CST = "9.1"
}
```

Listing 1: An example BibTeX entry retrieved from Science Direct for the article of Steinmacher et al. [46]. The additional entries we extract using APIs are highlighted in blue.

to aggregate data from various sources and deliver statistics in an organized manner, for example, Plum Analytics,[2] Altmetrics Explorer,[3] and ImpactStory.[4]

3 PROPOSED TECHNIQUE

In the following, we describe our technique, for which we depict an overview in Figure 2.

3.1 Premises

Our technique is concerned solely with facilitating two activities of conducting a literature analysis (cf. Figure 1), with our implementation missing any integration with other techniques to support, for example, the planning phase or actual search. Thus, before our technique can be used, the analyst has to complete the planning phase of their literature analysis, namely defining research questions, the search strategy, and selection criteria. In addition, the actual search must be performed, so that a list of potentially relevant publications is available as a BibTeX file.

When designing our technique (e.g., file formats, available information), we focused on four established digital libraries in software engineering, namely the ACM Digital Library,[5] IEEE Xplore,[6] Scopus,[7] and Science Direct.[8] These digital libraries allow analysts to download a collection of their search results as a BibTeX file (among others) that comprises most information we require (cf. Section 3.2).

[2] http://www.plumanalytics.com/about.html
[3] http://altmetric.com/
[4] http://impactstory.org/
[5] https://dl.acm.org/
[6] https://ieeexplore.ieee.org
[7] https://www.scopus.com
[8] https://www.sciencedirect.com/

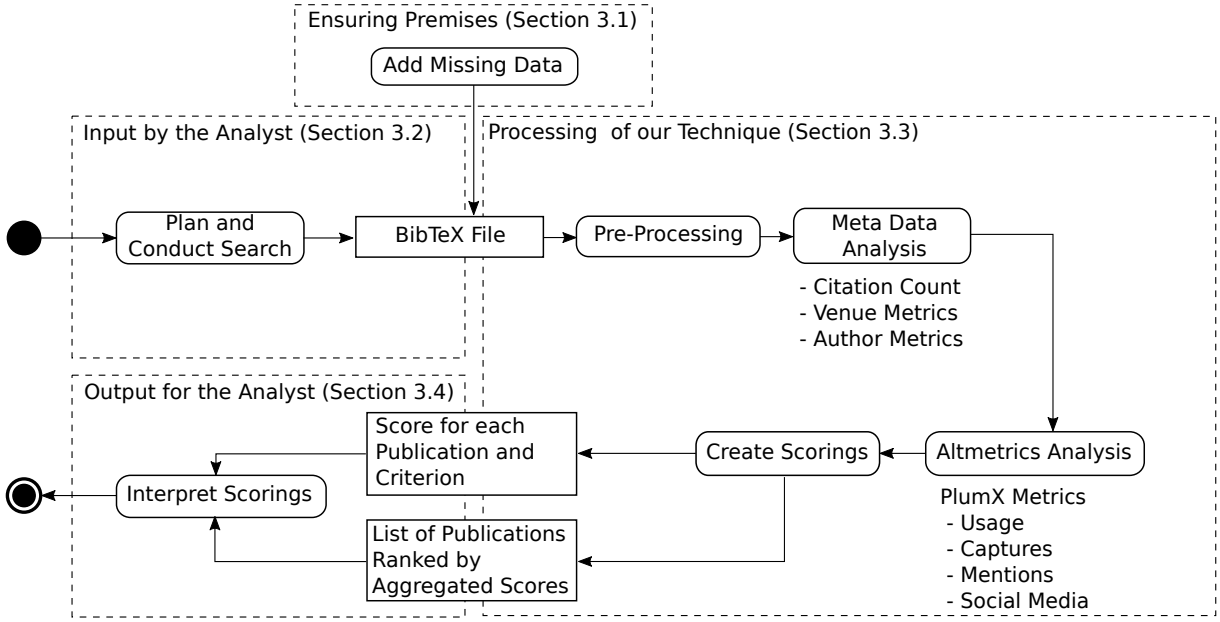

Figure 2: Overview of our metrics-based technique.

For our current prototype, we decided to rely on the format provided by Scopus and Science Direct (cf. Listing 1). Still, integrating other digital libraries is straightforward, either by transforming the format of the files or implementing a new parser for a library.

Some metrics that we use for our technique are supported by most digital libraries, but are usually not included in the BibTeX files. To solve this problem, we use different APIs to automatically extract citation counts and missing venue metrics (exemplified in blue in Listing 1). For instance, we retrieve citations through the Abstract Citations Count API,[9] that builds on Elsevier and Scopus databases. The journal metrics are supported exclusively by Scopus, and we extract them through the Serial Title API.[10] So, we use APIs to automate the retrieval of missing information, but this can be adopted to the analyst's needs—either by providing the data manually or by adopting our implementation to other sources. We decided to explicitly not enforce specific sources since metrics, especially citation counts, can vary heavily between sources (e.g., Google Scholar or the ACM Digital Library) depending on the publications these index. Note that the availability of metrics through APIs for *all* retrieved publications was our main reason for focusing on Scopus and Elsevier for now.

3.2 Input

Our technique requires the analyst to plan and conduct the search to then build a dataset from the results for the remaining analyses of our technique. This dataset is stored as the aforementioned BibTeX file of potentially relevant publications obtained, for example, through automatic searches in digital libraries. This BibTeX file must comprise the following fields and values (see Listing 1 for an example): author, title, abstract, keywords, venue, year, Digital Object Identifier (DOI), citation count, SCImago Journal Rank (SJR),[11] Source Normalized Impact per Paper (SNIP) [30], CiteScore,[12] and CiteScore Tracker (CST).

In short, the provided BibTeX file represents the data sample to be processed by our technique. Since, our technique relies on the dataset provided, the analyst must ensure completeness, for instance, through manual searches and snowballing, and combine the results to fulfill the aforementioned requirements. It is possible to use multiple search strings, for example, if literature from multiple domains is under investigation. In such a case, the analysts must provide all the BibTeX files to be processed independently.

3.3 Processing

Our technique builds on three sub-steps when processing the BibTeX file: (1) pre-processing the provided inputs for the analysis; (2) analyzing the meta data; and (3) collecting as well as interpreting altmetrics. In the following, we describe the details of each step, during which we consider different properties of a publication, namely its authors, venue, scientific impact, and usage.

3.3.1 Pre-Processing. Before performing the actual analysis, our technique parses the provided BibTeX file (cf. Listing 1) and inserts it into a database, building on the field names (occurring after a comma) and their values (in quotations).

3.3.2 Analyzing Meta Data. For assessing a publication, we examine the meta data comprised in the BibTeX file (cf. Listing 1). To this end, we define two assessment criteria (ACs) based on the data that is publicly available in digital libraries or databases. In Table 1, we provide an overview of these criteria, along with the corresponding

[9] https://dev.elsevier.com/cited_by_scopus.html
[10] https://dev.elsevier.com/documentation/SerialTitleAPI.wadl
[11] https://www.scimagojr.com/help.php
[12] https://journalmetrics.scopus.com/index.php/Faqs

meta data we use to assign scores to individual metrics that contribute to each of them. Next, we describe the use and reason about the usefulness of each criterion and its metrics.

AC_1 *How valuable is the publication in the context of the overall research being performed?*
Observing the citation relationships of a publication can help determine its importance within the research community, representing a useful indicator of the significance and impact of that publication. Additionally, altmetrics provide immediate and useful insights into the impact of publications and indicate their popularity. Overall, we combine these different metrics to assign a significance score to each publication.

Citation Count. Publications that provide useful and valid findings are usually referred and cited by other researchers. Thus, the citation counts provided by digital libraries are often considered as indicators for importance, visibility, and the overall impact of publications [3, 44]. Although this metric is conveniently available online (i.e., we rely on the Scopus database for extracting the citation counts), researchers are particularly concerned about its credibility as a quality indicator, mainly due to self-citations and the possibility of bias [19, 39]. Still, there is adequate evidence showing the usefulness of citation counts for assessing publications [6, 25, 26]. So, this indicator is widely supported and used to determine various sub-metrics, including average citations per year, the h-index of authors, Field-Weighted Citation Impact (FWCImpact) of publications, and CiteScore for publishing venues. We consider the citation count as an important quantitative metric that can help an analyst to assess quality in terms of the impact of a publication. In our technique, we normalize the citations between publications (using Equation 1 in Section 3.4) to allow for an easier comparison and for deriving an overall score. Note that we also use altmetrics for this assessment criterion, which we describe shortly to align with the analysis flow of our technique.

AC_2 *How credible are the findings of the performed research within its specific area?*
To assess the credibility of a publication, it can be helpful to determine the expertise of the authors and where the publication has been published. For this purpose, we combine author appearances and venue metrics to obtain scores reflecting the influence of a publication.

Author Contribution. For assessing a publication, it is meaningful to consider the authors and perceive their level of expertise. More contributions of an author to a specific research area clearly indicate a higher level of expertise. Analysts performing a literature analysis are especially considering the popularity of authors and give more importance to recognized authors, since their findings are likely to be relevant and valid. However, the contribution years must also be taken into account, as researchers actively contributing for a longer time have higher chances of being recognized by others. A useful metric in this regard is the h-index [9], which measures the productivity and citation impact of an individual—and which is supported particularly by Google Scholar and Scopus. However, due to access restrictions imposed by Scopus, our technique can currently only retrieve citation data, but not the h-index itself. To overcome this problem, we focus solely on the provided BibTeX file (i.e., the investigated research area) and use the citation scores of the included publications to define an h-index-like value. Precisely, we use the same formula, but only on the provided subset of an author's publications, which may be more representative of the specific research area. Afterwards, we add up and normalize the values to assign the resulting score to each publication. We remark that this can rate multiple authors with a smaller h-index higher than an author with a high h-index, but we kept this property, arguing that this accounts for additional validation by co-authors.

Publication Venue. Digital libraries, particularly Scopus, provide several venue metrics, allowing researchers to obtain an understanding of their perceived importance. Of these, we include:

- *SJR*: A quantitative value representing the average number of weighted citations a publication received during a selected year by publications published at the same venue during the previous three years.
- *SNIP*: A measure of the contextual citation impact determined by weighting individual citations based on the total number of citations in a research area, aiming to incorporate different citation practices between research areas.
- *CiteScore*: A metric for determining the impact of venues by measuring the average citations per publication over a three-year period to consistently observe the performance of a venue. Actually, the CiteScore metrics suite comprises eight different indicators: CiteScore itself, CiteScore Tracker, CiteScore Percentile, CiteScore Quartiles, CiteScore Rank, Citation Count, Document Count, and Percentage Cited. However, out of these, we only include the CiteScore Tracker (CST) as an additional venue metric, which is computed similarly as CiteScore, but for the current year only.

Although, there are certain web-portals, such as Core,[13] Conference Ranks[14] or Guide2Research[15] that provide different conference rankings, we currently rely only on the above data, which we extract from Scopus.

3.3.3 Using Altmetrics (PlumX). For our technique (i.e., AC_1), we use Plum Analytics' PlumX tool to retrieve altmetrics. PlumX provides one of the most detailed altmetrics-based datasets gathered from a variety of sources [10, 17], and it is integrated with Scopus, which is one of the major sources of scientific and technical publications. PlumX metrics consist of the following five categories;

- *Citations* contain both, citation counts from Scopus as well as other sources, such as, CrossRef, PubMed, and SciELO.
- *Usage* is a publication statistic summarizing several values, such as, abstract views, full-text downloads, and the number of URL clicks.
- *Mentions* indicate how often other people engage with a publication through blogs, comments, and reviews.
- *Social Media* provides values indicating the interaction on platforms, such as, Facebook, Twitter, Amazon, and Youtube, based on the number of likes, tweets, and shares.
- *Captures* track the interest of the audience based on, for example, number of readers, bookmarks, and citation exports.

[13] http://www.core.edu.au/conference-portal
[14] http://www.conferenceranks.com/
[15] http://www.guide2research.com/topconf/

Table 1: Overview of the assessment criteria we address by analyzing bibliographic data.

AC	Concrete Question	Data	Scoring
AC_1	How valuable is the publication in the context of the overall research being performed?	Citation Count	Normalized [0,1]
		Altmetrics	Normalized [0,1]
AC_2	How credible are the findings of the performed research within its specific area?	Author contribution	Normalized [0,1]
		Publication venue	Normalized [0,1]

We extract these altmetrics through the PlumX Metrics API for Scopus, building API requests based on the DOIs specified in the BibTeX file. Then, we normalize the retrieved values of each category to assign individual scores between 0 and 1. Also, we assign 0 in case of missing PlumX data, since these have only been introduced recently, and publications published before that point in time lack such data. However, we argue that altmetrics are not only emerging, but also a useful means of information, providing an understanding of what publications are relevant and of high quality before they obtain citations and broad visibility. We will discuss this assumption in more detail within Section 4.

3.4 Output

Based on the defined metrics we discussed in Section 3.3, our technique assigns an individual score to each publication. We score the metrics using a normalization strategy, which we display in Table 1. Precisely, we normalize values to obtain scores between 0 and 1, which is mostly the case for any continuously measurable metric. We use Equation 1 to normalize the values, where we divide the original value, such as citation count, with the maximum value we observe for the publications in the provided BibTeX file.

$$Score_{norm}(publication_i) = \frac{value(publication_i)}{value_{max}(publications)} \quad (1)$$

Finally, we accumulate all individual scores for each metric, using Equation 2, to assign a final score for every publication.

$$FinalScore(publication_i) = \sum_{k=1}^{k=14} Score_k(publication_i) \quad (2)$$

Then, our technique provides a list of all publications with their corresponding scores (i.e., individually for each metric and final for the publication) to the analyst in descending order of final scores.

4 EVALUATION

To conduct an evaluation, we implemented our technique as a prototype in Python, and used it to replicate 10 existing, manually performed systematic literature reviews and systematic mapping studies. In Table 2, we list the 10 reviews that we selected randomly from different research topics within the computer-science domain. As requirements for inclusion, we ensured that each of the selected reviews has a sufficient explanation of the employed research method. Particularly, the search strings for the automated search across digital libraries and the complete search strategy with the included data sources as well as inclusion and exclusion criteria must be reported. Furthermore, the number of publications retrieved from each data source must be mentioned along with the ones selected as relevant to allow us to compare the results of the original search with our replication. Lastly, we selected few reviews that are comparably old. Instead, we focused on more recent reviews to also evaluate the impact of altmetrics, which do rarely exist for older publications.

For each selected review, we employed the same search procedure as explained by the authors. Since we replicated the searches at a different point in time, the returned publications are not identical, as we display in Table 2. We can clearly observe the differences in the original and replicated search results, which are mainly caused by changes and technical issues of digital libraries [22, 24, 42]. To ensure that all relevant papers are included, analysts can perform a separate snowballing and add the newly found papers manually in the BibTeX file. However, for our initial experiment, we only considered the results from the automated searches reported in the original reviews. After employing the automated search, we downloaded and merged the BibTeX files provided by the digital libraries, using the result as input for our prototype. Note that our prototype currently only supports Scopus and Science Direct, which is why we only considered these sources.

To not rely on our interpretation of what publications are relevant and of what quality, which will differ from the original authors and misses their expertise of the research area, we decided to employ a fully automated, and thus more pessimistic evaluation. More precisely, we determined the validity of our prototype's scorings based on confusion matrices to measure its accuracy, precision, recall, and F1-score. To improve the comprehensibility of our study design and provide qualitative insights, we first exemplify our methodology for one of the reviews before presenting the overall outcomes.

4.1 Evaluation Example

We selected the systematic literature review of Nuñez et al. [32] as example. The review was originally conducted to review techniques for improving web accessibility, the domains these cover, and the disabilities that they address, covering the period from 2015 to 2018. Building on the guidelines of Kitchenham and Charters [22], the review describes four stages (the numbers relate to Figure 1): (1) specifying research questions, (4) implementing a search strategy, (5) selecting publications, and (7) synthesizing relevant information. We used the descriptions for the first three stages as follows.

Replicating the Search. To obtain relevant publications for their study, Nuñez et al. defined the following search string;

```
"('web application*' OR 'website*' OR 'web page*')
AND ('web accessibility')
AND ('method*' OR 'technique*')
AND ('evaluation' OR 'verification' OR 'validation')"
```

Table 2: Reviews we selected to evaluate our technique.

ID	Reference	Type	Year	Publications			
				Original		Replication	
				Found	Included	Found	Included
1	Turner et al. [48]	SLR	2009	256	27	211	23
2	Liu et al. [27]	SLR	2009	191	4	278	4
3	Steinmacher et al. [46]	SLR	2014	132	8	105	8
4	Knutas et al. [23]	SM	2015	88	24	70	18
5	Behutiye et al. [7]	SLR	2016	76	31	245	20
6	Baqais and Alshayeb [5]	SLR	2019	181	27	172	9
7	Al-Shaaby et al. [4]	SLR	2019	335	13	152	12
8	García-García et al. [18]	SLR	2019	293	6	99	6
9	Silva et al. [45]	SLR	2019	90	16	92	11
10	Nuñez et al. [32]	SM	2019	43	7	57	5

SLR: systematic literature review; SM: systematic mapping study.

Originally, Nuñez et al. performed an automated search across several digital libraries, including the ACM Digital Library, IEEE Xplore, Springer Link, and Scopus. For our analysis, we focused on Scopus (and Science Direct for other reviews), most importantly because PlumX data is specifically supported by this digital library, and our prototype is currently optimized for the data format of Scopus (e.g., BibTeX format, available APIs). As we can see in Table 2, Nuñez et al. received 43 search results from Scopus and selected seven publications to be relevant. In our replication, we received 57 results (14 more than the original review).

Employing the Prototype. After replicating the original search, we instantiated our prototype with the obtained BibTeX file, which contains the publications' meta data and the automatically recovered PlumX data. In Table 3, we display the resulting scorings our prototype assigned for each metric we defined. We can see that our replicated search comprised only five out of the seven publications selected by Nuñez et al..

Interpreting the Results. The scores we obtained (for *citation count*, *author contribution*, and *venue metrics*) and display in Table 3 are assigned based on a publication's meta data, providing information regarding its impact. Considering *citation counts*, the average final score for all 57 publications identified is 0.20. We can see that except for one publication, all relevant publications are above that value, with publication 17 being considerably higher scored. This may be closely related to the higher author contribution score that, combined with the citation metrics, may imply that the publication has been published by renowned experts of the research topic who have a high visibility. In general, the closer the citation count is to 1, the more often a publication has been cited in relation to the other publications identified.

The next score, for *author contribution*, in this example indicates that few relevant publications are published by the same authors within the specific research area. We remark that our technique currently supports author analyses based on Scopus only, and thus publications are missing from our data set. Furthermore, we show the four different *venue metrics* we implemented in our technique. We can see that three of the relevant publications receive comparably high score, indicating that they have been published at recognized venues compared to the other publications. Still, such metrics are not available for all publications. For this reason, the bottom two publications in Table 3 received 0 scores, which also reduced their final score.

Finally, we display the individual *PlumX* metrics. During our evaluation, we observed that this information is still missing for many papers, particularly those from earlier years. We understand that altmetrics are a rather new concept and accumulating this information for older papers would be challenging. However, due to their increasing popularity and support by digital libraries, we believe they are a useful means for indicating the importance of a publication. Our results in Table 3 show that most of the studies lack the PlumX data, exceptions being the capture count reflecting the number of readers interested in a publication.

To obtain the final score, we first calculate the average score over all 57 publications. On average, each publication received a final score of 0.40, and we can see that three of the relevant publications are above this average (i.e., 2, 21, 17). In contrast, of the remaining two publications one is closely below the average (41), while the lowest score is assigned to publication 16 (i.e., 0.09). To evaluate our technique, we use our data to construct confusion matrices: Publications with final scores above the average that were also included in the original review are true positives. The remaining publications with a higher score than the average are counted as false positives. Publications that are assigned lower final scores than the average are false negatives, while the rest are considered true negatives.

In Table 4, we display a summary of the confusion matrices and performance measurements of our evaluation. We can see that for the review of Nuñez et al. (i.e., 10), our technique achieves an accuracy and recall of 60 %. However, precision (13 %) and F1-score (21 %) are rather small. We expected such results, since automated searches for literature reviews usually result in a large number of irrelevant publications.

Table 3: Scores of the publications from the replicated search that were also included by Nuñez et al. [32].

Original ID	Citation Count	Author Contribution	Venue: SJR	Venue: SNIP	Venue: CiteScore	Venue: CST	PlumX: Usage	PlumX: Mentions	PlumX: Social Media	PlumX: Captures	Final Score
2	0.29	0.05	0.56	0.55	0.36	0.36	0.11	0.00	0.33	0.37	3.00
21	0.24	0.00	0.29	0.42	0.18	0.17	0.00	0.00	0.33	0.11	1.74
17	0.47	0.20	0.12	0.24	0.09	0.08	0.00	0.00	0.00	0.10	1.29
41	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.33
16	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.09

4.2 Overall Results

We further display the overall results of our evaluation in Table 4. Using our technique, we achieved an average accuracy of 53 % with a recall of 71 %. The average precision and F1-scores are relatively low with 14 % and 21 %, respectively. Considering the high number of false positives, this can be expected since our technique is subject to specifically defined data (i.e., meta and PlumX data) for the analysis, aiming to summarize the most relevant publications. Particularly, since fully automating the selection and assessment of publications to entirely replace manual analyses is impossible and the initial searches of systematic literature analyses often include a high number of irrelevant publications, we expected such numbers and argue that the scorings provide a good indication of what publications may be relevant and of what impact. So, our technique shows that a pure metrics-based analysis of publications is, not surprisingly, limited in its capabilities. However, it can support researchers by reducing the time needed to analyze publications, reason about the overall importance of the found publications based on bibliometrics, and get an overview of publications that may be unavailable to them. Interestingly, the performance of our technique improves for more recent studies (i.e., 4–10). This seems to be caused by recent publications having more accurate meta as well as PlumX data, which indicates the value of such metrics for selecting and assessing publications. It also indicates that our initial assumption regarding the value of altmetrics to represent qualitative publications with currently missing impact is reasonable.

5 DISCUSSION

In this section, we discuss the results of our evaluation in more detail, aiming to reason about the feasibility of our technique. To this end, we compare our technique with other techniques that employ similar ideas. Moreover, we reflect on the overall usability and limitations of our technique.

5.1 Comparison with other Techniques

Among existing techniques [29, 43], we found that PaperQuest [36] is the only one following a similar idea as we are. Precisely, PaperQuest analyzes citation links to identify and assess publications for a literature review. For this purpose, an analyst provides seed publications that are used to automatically build a dataset by utilizing the citation counts of an external library, such as Google Scholar. PaperQuest supports the analyst with efficient reading decisions, sorting publications according to the ones they found interesting. Based on the analyst's selection along with the citation counts of the publications, PaperQuest continues to update the reading list.

Since our technique uses a different analysis, metrics, and workflow than PaperQuest, a direct comparison can only provide hints on the feasibility of our technique. Nonetheless, the core idea of using meta data, such as citation counts, for (semi-)automating the selection process is identical for both techniques. Even though PaperQuest is still in its early development stage and lacks formal evaluation, the developers gathered preliminary feedback from students and faculty members. The results show that PaperQuest reduces the effort required for conducting a literature analysis by supporting researchers in making their decision regarding a publication. While no actual measurements are provided, this indicates that our technique can also be a highly valuable means.

Other studies compared the work load and recall of replicating systematic literature reviews with other techniques (e.g., for visual text mining) and analysts. Compared to these studies, we did not involve any analyst, but only employed our metrics to evaluate the feasibility of our technique. Interestingly, different techniques and workflows can heavily vary in their performance. For instance, Abilio et al. [1] replicated literature reviews with two different strategies (i.e., vector models and ranking functions over search strings), which yielded between 17.2 % and 52 % precision with 80 % and 90 % recall. Other studies [34, 38] support these findings, usually indicating a high recall with varying error rates (precision) when supporting analysts with semi-automated techniques. Considering that these studies involve humans replicating the conduct, we argue that the performance of our technique based on a semi-automated analysis is highly promising. Only relying on metrics, we achieve a similar recall with a precision at the lower boundary. So, we argue that our technique is valuable for selecting and assessing publications, and can help to reduce the required manual effort.

Table 4: Summary of the evaluation results: confusion matrix and performance measures.

ID	Confusion Matrix				Performance Measures			
	TP	FN	FP	TN	Accuracy	Precision	Recall	F1-Score
1	20	3	104	84	0.49	0.16	0.87	0.27
2	4	0	147	127	0.51	0.03	1.00	0.06
3	5	3	58	39	0.42	0.08	0.63	0.15
4	14	4	16	36	0.71	0.47	0.78	0.59
5	12	8	159	66	0.32	0.07	0.60	0.13
6	9	0	75	88	0.56	0.11	1.00	0.20
7	2	10	62	78	0.53	0.03	0.17	0.05
8	4	2	43	50	0.55	0.09	0.67	0.16
9	9	2	36	45	0.59	0.20	0.82	0.32
10	3	2	21	31	0.60	0.13	0.60	0.21
			Average:		0.53	0.14	0.71	0.21

TP: true positive; TN: true negative; FP: false positive; FN: false negative.

5.2 Limitations and Threats to Validity

Our technique is limited to the analysis of meta data as well as altmetrics, which cannot replace a critical, manual assessment of publications. To ease this process, our technique successfully indicates a large amount of irrelevant publications automatically, allowing an analyst to focus on the most relevant and promising ones. Still, our technique also faces technical limitations. For example, at the moment, we support only two digital libraries to their full extent and are aware of some bugs when parsing special characters in BibTeX files.

Besides such technical limitations, our evaluation also face some threats to validity. Namely, we did replicate only 10 literature reviews with a minimum of human involvement. However, the semi-automated scoring performed well compared to similar tools, and we argue that the final scores would improve far more when involving human subjects critically analyzing the automated decisions. A major advantage of our evaluation is the fact that we avoid several other threats and problems (e.g., subjective opinions on relevancy, obtaining domain knowledge), which is why we decided for using the described performance measurements.

6 CONCLUSION

In this paper, we proposed a technique for facilitating the selection and assessment of publications during literature analyses. For this purpose, we rely on meta data and altmetrics, which allow us to derive a rating based on publicly available data. We argue that our technique allows an analyst to better understand the relevance and importance of identified publications by providing such ratings. Employing our technique in an evaluation with 10 replicated literature analyses, we obtained comparable results to other techniques (i.e., accuracy of 53 %, recall of 71 %, precision of 14 %, and F1-score of 21 %). Other techniques are typically evaluated based on fewer replications with human subjects. So, we assume that the comparable results of our technique represent a rather pessimistic scenario (i.e., involving a human subject would improve the results). Overall, we argue that our technique and its comprised metrics are a complementary and helpful means to facilitate literature analyses—and cannot replace a careful evaluation by an expert.

In future work, we aim to extend our prototypical implementation regarding two dimensions. First, we aim to improve the selection and assessment by improving our technique and incorporating other concepts, such as content analysis. Second, we intend to cover all phases of a systematic literature review, enabling researchers to cover all phases within a single technique.

REFERENCES

[1] Ramon Abilio, Flávio Morais, Gustavo Vale, Claudiane Oliveira, Denilson Pereira, and Heitor Costa. 2015. Applying Information Retrieval Techniques to Detect Duplicates and to Rank References in the Preliminary Phases of Systematic Literature Reviews. *CLEI Electronic Journal* 18, 2 (2015), 1–24.

[2] Kuku J. Aduku, Mike Thelwall, and Kayvan Kousha. 2017. Do Mendeley Reader Counts Reflect the Scholarly Impact of Conference Papers? An Investigation of Computer Science and Engineering. *Scientometrics* 112, 1 (2017), 573–581.

[3] Dag W. Aksnes, Liv Langfeldt, and Paul Wouters. 2019. Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories. *SAGE Open* 9, 1 (2019).

[4] Ahmed Al-Shaaby, Hamoud Aljamaan, and Mohammad Alshayeb. 2020. Bad Smell Detection Using Machine Learning Techniques: A Systematic Literature Review. *Arabian Journal for Science and Engineering* (2020), 1–29.

[5] Abdulrahman A. B. Baqais and Mohammad Alshayeb. 2019. Automatic Software Refactoring: A Systematic Literature Review. *Software Quality Journal* (2019), 1–44.

[6] Joeran Beel and Bela Gipp. 2009. Google Scholar's Ranking Algorithm: The Impact of Citation Counts (An Empirical Study). In *International Conference on Research Challenges in Information Systems (RCIS)*. 439–446.

[7] Woubshet N. Behutiye, Pilar Rodríguez, Markku Oivo, and Ayşe Tosun. 2017. Analyzing the Concept of Technical Debt in the Context of Agile Software Development: A Systematic Literature Review. *Information and Software Technology* 82 (2017), 139–158.

[8] Lutz Bornmann. 2014. Do Altmetrics Point to the Broader Impact of Research? An Overview of Benefits and Disadvantages of Altmetrics. *Journal of Informetrics* 8, 4 (2014), 895–903.

[9] Lutz Bornmann and Hans-Dieter Daniel. 2007. What do we Know About the h-Index? *Journal of the Association for Information Science and Technology* 58, 9 (2007), 1381–1385.

[10] Tara J. Brigham. 2014. An Introduction to Altmetrics. *Medical Reference Services Quarterly* 33, 4 (2014), 438–447.

[11] Meredith Brown. 2014. Is Almetrics an Acceptable Replacement for Citation Counts and the Impact Factor? *The Serials Librarian* 67, 1 (2014), 27–30.

[12] Rodrigo Costas, Zohreh Zahedi, and Paul Wouters. 2015. Do "Altmetrics" Correlate with Citations? Extensive Comparison of Altmetric Indicators with Citations from a Multidisciplinary Perspective. *Journal of the Association for Information Science and Technology* 66, 10 (2015), 2003–2019.

[13] David Crotty. 2014. Altmetrics: Finding Meaningful Needles in the Data Haystack. *Serials review* 40, 3 (2014), 141–146.

[14] Tore Dybå and Torgeir Dingsøyr. 2008. Strength of Evidence in Systematic Reviews in Software Engineering. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*. 178–187.

[15] Gunther Eysenbach. 2011. Can Tweets Predict Citations? Metrics of Social Impact based on Twitter and Correlation with Traditional Metrics of Scientific Impact. *Journal of Medical Internet Research* 13, 4 (2011).

[16] Sandra Fabbri, Cleiton Silva, Elis Hernandes, Fábio Octaviano, André D. Thommazo, and Anderson Belgamo. 2016. Improvements in the StArt Tool to Better Support the Systematic Review Process. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*. ACM, 1–5.

[17] Linda M. Galloway, Janet L. Pease, and Anne E. Rauh. 2013. Introduction to Altmetrics for Science, Technology, Engineering, and Mathematics (STEM) Librarians. *Science & Technology Libraries* 32, 4 (2013), 335–345.

[18] Julián A. García-García, Jose G. Enríquez, Mercedes Ruiz, Carlos Arevalo, and Andrés Jiménez-Ramírez. 2020. Software Process Simulation Modelling: Systematic Literature Review. *Computer Standards & Interfaces* (2020).

[19] Ioannis A. Giannakakis, Anna-Bettina Haidich, Despina G. Contopoulos-Ioannidis, George N. Papanikolaou, Maria S. Baltogianni, and John P. A. Ioannidis. 2002. Citation of Randomized Evidence in Support of Guidelines of Therapeutic and Preventive Interventions. *Journal of Clinical Epidemiology* 55, 6 (2002), 545–555.

[20] Edgar Hassler, Jeffrey C. Carver, David Hale, and Ahmed Al-Zubidy. 2016. Identification of SLR Tool Needs – Results of a Community Workshop. *Information and Software Technology* 70 (2016), 122–129.

[21] Barbara A. Kitchenham, David Budgen, and Pearl Brereton. 2015. *Evidence-Based Software Engineering and Systematic Reviews*. CRC Press.

[22] Barbara A. Kitchenham and Stuart Charters. 2007. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. Technical Report EBSE-2007-01. Keele University and University of Durham.

[23] Antti Knutas, Jouni Ikonen, and Jari Porras. 2015. Computer-supported Collaborative Learning in Software Engineering Education: A Systematic Mapping Study. *International Journal on Information Technologies & Security* 7, 4 (2015), 45–72.

[24] Jacob Krüger, Christian Lausberger, Ivonne von Nostitz-Wallwitz, Gunter Saake, and Thomas Leich. 2019. Search. Review. Repeat? An Empirical Study of Threats to Replicating SLR Searches. *Empirical Software Engineering* (2019), 1–51.

[25] Abhaya V. Kulkarni, Brittany Aziz, Iffat Shams, and Jason W. Busse. 2009. Comparisons of Citations in Web of Science, Scopus, and Google Scholar for Articles Published in General Medical Journals. *JAMA: Journal of the American Medical Association* 302, 10 (2009), 1092–1096.

[26] Duncan Lindsey. 1989. Using Citation Counts as a Measure of Quality in Science Measuring What's Measurable Rather than What's Valid. *Scientometrics* 15, 3-4 (1989), 189–203.

[27] Dapeng Liu, Qing Wang, and Junchao Xiao. 2009. The Role of Software Process Simulation Modeling in Software Risk Management: A Systematic Review. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 302–311.

[28] Viviane Malheiros, Erika Höhnr, Roberto Pinho, Manoel Mendonca, and José C. Maldonado. 2007. A Visual Text Mining Approach for Systematic Reviews. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 245–254.

[29] Christopher Marshall and Pearl Brereton. 2013. Tools to Support Systematic Literature Reviews in Software Engineering: A Mapping Study. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 296–299.

[30] Henk F. Moed. 2010. Measuring Contextual Citation Impact of Scientific Journals. *Journal of Informetrics* 4, 3 (2010), 265–277.

[31] Ehsan Mohammadi, Mike Thelwall, Stefanie Haustein, and Vincent Larivière. 2015. Who Reads Research Articles? An Altmetrics Analysis of Mendeley User Categories. *Journal of the Association for Information Science and Technology* 66, 9 (2015), 1832–1846.

[32] Almendra Nuñez, Arturo Moquillaza, and Freddy Paz. 2019. Web Accessibility Evaluation Methods: A Systematic Review. In *Design, User Experience, and Usability. Practice and Case Studies*. Springer, 226–237.

[33] Andrea G. Nuzzolese, Paolo Ciancarini, Aldo Gangemi, Silvio Peroni, Francesco Poggi, and Valentina Presutti. 2019. Do Altmetrics Work for Assessing Research Quality? *Scientometrics* 118, 2 (2019), 539–562.

[34] Fábio R. Octaviano, Katia R. Felizardo, José C. Maldonado, and Sandra C. P. F. Fabbri. 2015. Semi-automatic Selection of Primary Studies in Systematic Literature Reviews: Is it Reasonable? *Empirical Software Engineering* 20, 6 (2015), 1898–1917.

[35] Guy Paré, Marie-Claude Trudel, Mirou Jaana, and Spyros Kitsiou. 2015. Synthesizing Information Systems Knowledge: A Typology of Literature Reviews. *Information & Management* 52, 2 (2015), 183–199.

[36] Antoine Ponsard, Francisco Escalona, and Tamara Munzner. 2016. PaperQuest: A Visualization Tool to Support Literature Review. In *Conference on Human Factors in Computing Systems (CHI)*. ACM, 2264–2271.

[37] Jason Priem, Paul Groth, and Dario Taraborelli. 2012. The Altmetrics Collection. *PLoS one* 7, 11 (2012).

[38] Giuseppe Rizzo, Federico Tomassetti, Antonio Vetro, Luca Ardito, Marco Torchiano, Maurizio Morisio, and Raphael Troncy. 2017. Semantic Enrichment for Recommendation of Primary Studies in a Systematic Literature Review. *Digital Scholarship in the Humanities* 32, 1 (2017), 195–208.

[39] Karen A. Robinson and Steven Goodman. 2011. A Systematic Examination of the Citation of Prior Research in Reports of Randomized, Controlled Trials. *Annals of Internal Medicine* 154, 1 (2011), 50–55.

[40] Yusra Shakeel, Rand Alchokr, Jacob Krüger, Gunter Saake, and Thomas Leich. 2021. Are Altmetrics Proxies or Complements to Citations for Assessing Impact in Computer Science?. In *Joint Conference on Digital Libraries (JCDL)*. IEEE, 284–286.

[41] Yusra Shakeel, Rand Alchokr, Jacob Krüger, Gunter Saake, and Thomas Leich. 2022. Altmetrics and Citation Counts: An Empirical Analysis of the Computer Science Domain. In *Joint Conference on Digital Libraries (JCDL)*. IEEE.

[42] Yusra Shakeel, Jacob Krüger, Ivonne von Nostitz-Wallwitz, Christian Lausberger, Gabriel C. Durand, Gunter Saake, and Thomas Leich. 2018. (Automated) Literature Analysis - Threats and Experiences. In *International Workshop on Software Engineering for Science (SE4Science)*. ACM, 20–27.

[43] Yusra Shakeel, Jacob Krüger, Ivonne Von Nostitz-Wallwitz, Gunter Saake, and Thomas Leich. 2019. Automated Selection and Quality Assessment of Primary Studies: A Systematic Literature Review. *Journal of Data and Information Quality* 12, 1 (2019), 1–26.

[44] Yusra Shakeel, Jacob Krüger, Gunter Saake, and Thomas Leich. 2018. Indicating Studies' Quality based on Open Data in Digital Libraries. In *International Conference on Business Information Systems (BIS)*. Springer, 579–590.

[45] Rômulo S. Silva, Artur M. Mol, and Lucila Ishitani. 2019. Virtual Reality for Older Users: A Systematic Literature Review. *International Journal of Virtual Reality* 19, 1 (2019), 11–25.

[46] Igor Steinmacher, Marco A. G. Silva, Marco A. Gerosa, and David F. Redmiles. 2015. A Systematic Literature Review on the Barriers Faced by Newcomers to Open Source Software Projects. *Information and Software Technology* 59 (2015), 67–85.

[47] Mike Thelwall and Tamara Nevill. 2018. Could Scientists Use Altmetric.com Scores to Predict Longer Term Citation Counts? *Journal of Informetrics* 12, 1 (2018), 237–248.

[48] Mark Turner, Barbara A. Kitchenham, Pearl Brereton, Stuart Charters, and David Budgen. 2010. Does the Technology Acceptance Model Predict Actual Use? A Systematic Literature Review. *Information and Software Technology* 52, 5 (2010), 463 – 479.

[49] Jane Webster and Richard T. Watson. 2002. Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly* 26, 2 (2002), xiii–xxiii.

[50] You Zhou, He Zhang, Xin Huang, Song Yang, Muhammad A. Babar, and Hao Tang. 2015. Quality Assessment of Systematic Reviews in Software Engineering: A Tertiary Study. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*. ACM, 1–14.