Scientometrics (2020) 125:1517–1540
https://doi.org/10.1007/s11192-020-03558-7

How accurate are policy document mentions? A first look at the role of altmetrics database

Houqiang Yu[1] · Xueting Cao[1] · Tingting Xiao[2] · Zhenyi Yang[1]

Received: 3 December 2019 / Published online: 17 June 2020

Abstract
Policy document mention is considered to indicate the significance and societal impact of scientific product. However, the accuracy of policy document altmetrics data needs to be evaluated to fully understand its strength and limitation. An in-depth coding analysis was conducted on sample policy documents records of Altmetric.com database. The sample consists of 2079 records from all 79 distinct policy document source platforms tracked by the database. Errors about mentioned publications in the policy documents (type A error) are found in 8% of the records, while errors about either the recorded policy documents or the mentioned publications in the altmetrics database (type B error) are found in 70% of the records. In type B error, policy document link error (5% of the records) could be attributable to the policy document website, transcription error (52% of the records) could be attributable to the third-party bibliographic data provider. These two categories of error are relatively minor and may have limited influence on altmetrics research and practices. False positive policy document mention (13% of the records), however, could be attributable to the Altmetric database and may diminish the validity of research based on the policy document altmetrics data. The underlying reasons remain to be further investigated. Considering the high complexity of extracting mentions of publications from various sources and formats of policy documents as well as its short history, Altmetric database has achieved excellent performance.

Keywords Altmetrics · Data quality · Policy document · Altmetrics database · Accuracy

Introduction

Altmetrics has offered non-traditional way of measuring the diverse impact of scientific product. Altmetrics indicators have been widely adopted by various stakeholders for showcasing the general attention and the potential impact of their work. Many previous studies focused on altmetrics data sources like Twitter (Mohammadi et al. 2018; Yu 2017),

✉ Houqiang Yu
yuhouq@yeah.net

[1] School of Economics and Management, Nanjing University of Science and Technology, Nanjing, China

[2] Nanjing Library, Nanjing, China

Mendeley (Aduku et al. 2017; Zahedi and Haustein 2018) and F1000 (Bornmann and Haunschild 2015). Lately, policy document is studied as a potential useful data source for indicating the societal impact of scientific product (Bornmann et al. 2016), because policy usage of scientific product is supposed to reflect the relationship between academic research and policy making. The policy document is defined very broadly by the Altmetric.com company and refers to any policy, guidance, or guidelines document from a governmental or non-governmental organization.[1] Compared with other altmetrics data sources, policy document is particularly of interest in twofold, i.e. it is target-oriented and it focuses on a relevant part of society. Wooldridge and King (2018) found that altmetric score is highly correlated with the peer review scores of societal impact and news and policy sources are likely to contain the most informative data to assess impact.

Despite the popularity of altmetrics, many fundamental questions are not answered among which data quality is identified as one of the major challenges (Haustein 2016). According to NISO (2016), accuracy, transparency and replicability are three essential dimensions of data quality. Altmetrics databases are the infrastructure of altmetrics research and services. They constantly collect, clean and store altmetrics data in large scale and provide the data in a systematic and usable way. Altmetric.com database is by far the most commonly used altmetrics database.

Data quality of citation databases

Good quality of bibliographic data is fundamental to bibliometrics research and application. Abundant studies have been dedicated to examine the data quality of citation databases. These studies may provide relevant implication for studies of altmetrics data quality. To sum up, there are two perspectives of studying the topic.

The first perspective is to directly check the accuracy of bibliographic data in the database via content analysis. Investigated databases are mainly multidisciplinary databases like Web of Science (WoS), Scopus and Google Scholar (GS). Data errors can be classified into two major categories (Buchanan 2006), i.e. author error and database mapping error. Author error refers to error that is generated by the cited author, for example, the error can occur in the publication title, publication year, volume number or pagination. Many studies have investigated the first category of data error and reported its percentage. The percentage varies according to the subject area and the scope of author error as defined by the researcher. Based on a large-scale study, Moed (2002) estimated the percentage of author error in SCI database to be 7%. Database mapping error refers to error that is generated by citation databases, for example, transcription error, or cited-article omitted from a cited-article list. However, only a few studies have looked at the second category of data error, because in the past it is difficult to determine the rate of clerical errors introduced in the citation database (Garfield 1974). Franceschini et al. (2014) analyzed the omitted citations and found the omitted-citation rate of Web of Science to be 6%. Many weird errors are discovered in Scopus (Franceschini et al. 2016a). The database mapping error is further classified into errors in the transcription of author names or article title, incomplete cited-article list, omitted cited-article list, wrong or missing DOI, and errors concerning online-first papers (Franceschini et al. 2016b).

[1] See https://help.altmetric.com/support/solutions/articles/6000060968-what-outputs-and-sources-does-altmetric-track.

The second perspective is to measure the accuracy of bibliographic data by comparing different bibliographic databases. Meho and Yang (2007) compared the data quality of three citation databases, i.e. WoS, Scopus and Google Scholar (GS) and found GS has low accuracy and thus low reliability. Calver et al. (2017) studied the discrepancy of searching result between Scopus, Google Scholar, WoS and WoSCC, and found that retrieval result using the same keyword is different for different databases. Therefore, to improve the consistency across different databases will help users reduce the cost of finding proper database and improve the retrieval efficiency. Donner (2017) studied the influence of inaccurate document type on the citation index and found that data of accurate and inaccurate document type would lead to different citation counts. In addition, several comprehensive studies (Archambault et al. 2009; Harzing and Alakangas 2016; Wang and Waltman 2016) had tapped into the data quality of Scopus, GS and WoS.

Meanwhile, the underlying reasons of the data error and corresponding improving strategies are discussed. It is suggested that authors should be more rigorous and self-disciplined in citing behavior for reducing data error (Zhao 2009). Also, editors should put more efforts to check the authenticity and validity of references (Chen 2014). In database level, dictionary database structure can be used to automatically check the data with programs to reduce human cost (Su 2001). Although not all discovered errors are corrected (Franceschini et al. 2016c), bibliographic databases like Web of Science has made use of these relevant studies to improve their data quality (Prins et al. 2016).

Data quality of altmetrics databases

Similar to the situation of citation indicators, data quality is of critical importance for establishing credibility of altmetrics indicators. Without careful examination of the data quality, the application of altmetrics databases will be severely criticized. Accuracy is above all the basis of data reliability.

Research of altmetrics data quality is still in preliminary stage. Most previous studies compare across different altmetrics aggregators to check the coverage of different altmetrics data sources as collected and maintained by these aggregators (Chamberlain 2013; Peters et al. 2014; Zahedi et al. 2014). It is found that factors such as timing, platform and publication identifier will influence the reported altmetrics data (Meschede and Siebenlist 2018). Even for the same set of papers, different major aggregators provide different metrics (Ortega 2018, 2019). The differences pose challenges to the reliability of altmetrics. Zahedi and Costas (2018) investigated the underlying reasons and found that a range of different methodological, technical and reporting choices have determined the final counts, yet no universe recommendations could be made for data aggregators and users because each practice has its own cons and pros.

As regards the direct checking of the accuracy or completeness of altmetrics data, Zahedi et al. (2014) used a small sample to compare the metadata of publications that were presented in Mendeley and indexed in Web of Science and found that journal and article titles are the most erroneous. Unlike citations, Mendeley reader count is fluctuate and may decrease (Bar-Ilan 2014).

Studies of policy document altmetrics

Policy document usage of academic research is strong evidence of societal impact. According to Newson et al. (2018), two parallel streams of research are contributing to provide

insights about the influence of research on policy, i.e. research impact assessments that starts from looking at the research, and research use assessments that starts from looking at the policy document. For research impact assessments, commonly used data sources are interviews, surveys, policy documents, focus groups and direct observation. Case study is the most popular method. Many policy assessment studies use forward tracing approaches to examine single policy document to corroborate claimed impacts. However, these qualitative studies are limited by the data scale and high cost.

Policy document altmetrics has provided the opportunity to measure the usage of scientific product in policy document in a more systematic way. Several studies have explored the usability of policy document altmetrics. Using policy document altmetrics data provided by Altmetric.com company, Bornmann et al. (2016) find that only 1.2% of climate change publications have at least one policy mention. In another study, Haunschild and Bornmann (2017) find that less than 0.5% of papers indexed in Web of Science are mentioned at least once in policy-related documents. The percentage is lower than expected, because Khazragui and Hudson (2015) argue that policy tends to be based upon a large body of work. Nevertheless, Tattersall and Carroll (2018) used policy document altmetrics data collected by Altmetric.com company to evaluate the research from University of Sheffield. It is found that altmetric explorer database has offered important and highly accessible data on the policy impact, but the data must be used with caution because data errors are found.

Research questions

While policy document altmetrics can indicate the societal impact of scientific product, the accuracy of policy document altmetrics data is still unclear. The study is dedicated to measure the accuracy of policy document altmetrics data in Altmetric.com database, in order to inform the potential limitation of research or application that are based on the Altmetric.com policy document data and provide reference for improving the data quality in the future. To be specific, the research questions are:

1. Is there error in policy document altmetrics data? If so, what types of errors may occur?
2. What is the distribution of different categories of errors in policy document altmetrics data?

The result is compared with that of traditional citation databases. Potential underlying reasons of the data errors are discussed.

Methodology

Data source

As one of the major altmetrics data aggregators, Altmetric.com company has started to collect policy document mentions of scientific product in January 2013.[2] Data were retrieved

[2] See https://help.altmetric.com/support/solutions/articles/6000136884-when-did-altmetric-start-tracking-attention-to-each-attention-source.

Table 1 Productivity distribution of policy document source platforms (Top 10)

R	Policy document source platform	N.P.	%	N.M.	%
1	World Health Organization	10,935	13.6	240,140	17.3
2	Analysis & Policy Observatory (APO)	9592	11.9	123,493	8.9
3	National Academies Press	7565	9.4	295,981	21.3
4	UK Government (GOV.UK)	7025	8.7	57,792	4.2
5	Centers for Disease Control and Prevention (CDC)	6122	7.6	157,443	11.3
6	National Bureau of Economic Research	4992	6.2	46,987	3.4
7	World Bank	4132	5.1	31,316	2.3
8	European Food Safety Authority	3405	4.2	20,829	1.5
9	Food and Agriculture Organization of the United Nations	2813	3.5	35,467	2.6
10	rijksoverheid.nl	2794	3.5	25,530	1.8
Subtotal (of top 10 source platforms)		59,375	73.7	1,034,978	74.4

*N.P. is number of policy documents of each source platform that are captured by Altmetric.com database, N.M is number of policy document mentions of scientific product

on June 8th, 2018. In total, there are 1,390,350 policy document mentions of 1,072,397 unique scientific product, from 80,476 unique policy documents that are released by 79 unique source platforms. The top 10 source platforms are listed in Table 1. The length of policy document and the number of scientific products mentioned by the policy document have shown great difference. For example, the longest policy document (entitled with *Dietary Reference Intakes: The Essential Guide to Nutrient Requirements*) has 1345 pages and mentioned 3731 unique scientific products.

Sampling strategy

To measure the accuracy of policy document mentions, we begin by analyzing the general production process of altmetrics data, as is illustrated in Fig. 1. This can help identify the positions where potential errors may happen. Compared with traditional bibliometrics data, the production of altmetrics data has two additional steps, i.e. production of altmetrics source data (e.g. policy document) by the altmetrics source platform (e.g. policy document website), and production of the altmetrics data in altmetrics database (e.g. Altmetric database) accomplished by altmetrics companies (e.g. Altmetric.com company). Because the quality issue of bibliometrics data has been well studied, in this study, we focus only on these two additional steps of producing altmetrics data. Both steps could potentially introduce data errors. Errors found in these two steps are defined as two categories of data error.

Type A error: the error found about mentioned publications in the policy documents, by using the metadata of source publications as standards to compare with.

Type B error: the error found about either the recorded policy documents or the mentioned publications in the altmetrics database, by using the policy documents downloaded from the source website and publications in the source journal as standards to compare with. It must be noted that type B errors are not necessarily produced by the altmetrics database because, for example, third-party data may be used and related errors should not be attributable to the altmetrics database.

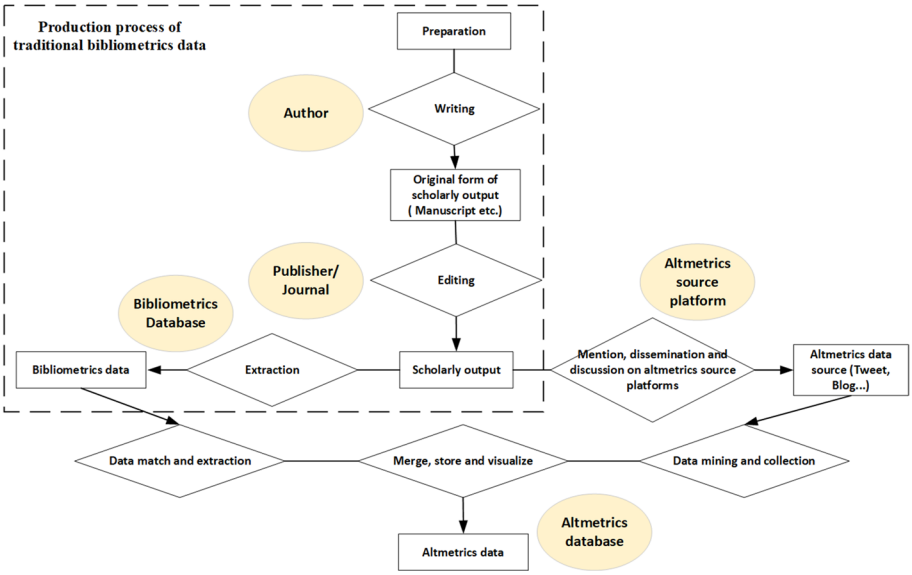


Fig. 1 Production process of altmetrics data

According to Altmetric.com company, each policy document source platform has its unique data structure and therefore they have created a personalized crawler for each source platform. Except for that, all records are supposed to share the same data extraction and matching technique that the company has used to automatically collect, merge and visualize altmetrics data. In total, there are 79 unique policy document source platforms. 30 policy documents from each source platform are randomly selected. For policy document source platform of which the total number of captured policy documents is no higher than 30, all policy documents are selected. For each policy document, one record of policy document mention is randomly selected. Each record has the relevant data of a policy document mentioning a scientific product. The final sample is a dataset of 2079 records.

Process of coding a single record

The process of coding a single record is demonstrated in Fig. 2. In general, there are three steps as described below.

1. Click the URL of policy document that is collected and recorded by the Altmetric.com database. If the URL doesn't work, the title of the policy document is used to search for the policy document, in order to examine the existence of such policy document.
2. After we find the original file of the policy document (mostly in pdf format), we download and open the file to search for the mentioned scientific product as recorded by the Altmetric.com database, in order to examine whether the scientific product is indeed mentioned in the policy document.
3. After we find the location of scientific product (mostly papers) mentioned in the policy document, we cross check the bibliographic information of the scientific product from three sources, i.e. the policy document file, the Altmetric.com detail page and the origi-

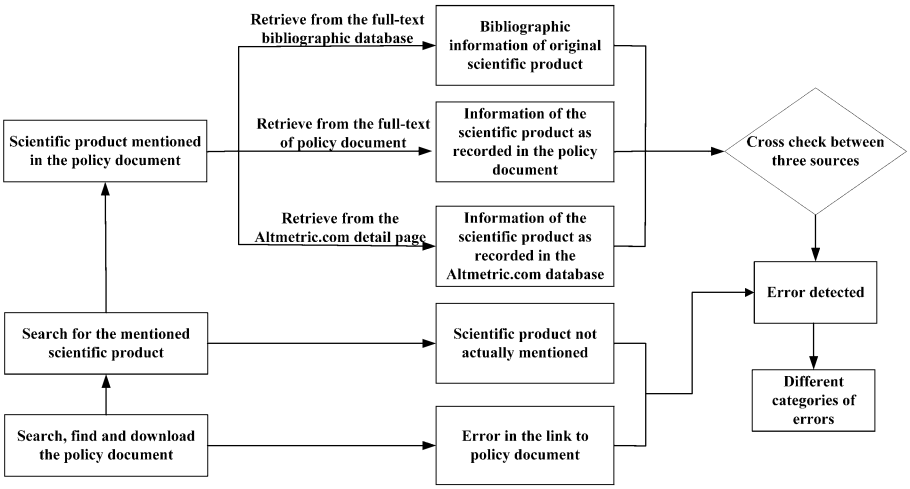


Fig. 2 Process of coding a single record of policy document mention

nal file of scientific product. The original file of scientific product is retrieved from full-text databases or open access repositories, and is used as benchmark with which any inconsistency is deemed as error. The error is further classified into different categories. For a certain number of records, two categories of errors could simultaneously occur and are coded. Very few records have more than two categories of errors, only two categories of errors that are supposed to be more important are coded.

Process of coding the whole sample dataset

The coding table is established via three steps.

1. 100 records are randomly selected from the sample dataset and are coded by four coders independently. The aim is twofold, to define the possible categories of errors and to standardize the coding process. Based on the coding practice, the initial coding table is created.
2. Another 100 records are randomly selected from the sample dataset and are coded by the four coders using the initial coding table and the standardized coding process. The aim is also twofold, to improve the usability of the coding table and to calculate the coding consistency. A few new categories of error are added to the initial coding table. The rate of coding consistency between the four coders is 86%. The records of coding inconsistency are discussed to understand the underlying reason. Based on the discussion and new discovery of categories of error, the final coding table is obtained (Table 6 and Table 7 in the "Appendix").

   Type A error is discovered by comparing the original file of scientific product and the information recorded in the policy document. Type B error is discovered by comparing the original file of scientific product, the policy document file and the Altmetric.com detail page.

Table 2 Distribution of the first level code data errors

First level code	Number of incorrect records	%
1 Type A error	169	8.1
2 Type B error	1445	69.5
Total	1523	73.2

*Total number of records is 2079

Table 3 Distribution of the second level code data errors (type A error)

Code	Definition	% of total incorrect records (%)
1.1 Author error of scientific product	107	5.1
1.2 Title error of scientific product	31	1.5
1.3 Publication date error of scientific product	19	0.9
1.4 Source (Journal) error of scientific product	12	0.6
Total	169	8.1

3. the total sample dataset of 2079 records are coded by two coders using the final coding table. The rate of coding consistency between the two coders is 98%. Records of inconsistent coding are discussed and coded with the help of a third coder.

Result

The general distribution of data errors is shown in Table 2. For the investigated dataset, 73% of records have data errors. The percentage is substantially high. 8% of records have type A error while 70% of records have type B error. 4% of records have both type A error and type B error.

Type A error

Table 3 shows the distribution of type A error. As shown in Table 3, code 1.1 error has the highest percentage. In every twenty policy document mentions of scientific product, one mention will wrongly record the author's name of the mentioned scientific product.

(Code 1.1) Author error of scientific product

A relatively frequent type A error concerns the wrong/missing author's name of mentioned publication in the policy document. The author's name of mentioned publication may be misspelled in several ways. Figure 3 (in the "Appendix") exemplifies that the author "Seager M." is misspelled as "Seagar M." in the policy document. In other cases, author whose name is not in English but in other language such as German or Russian may be spelled wrong. In a few cases, author's middle name is omitted.

Table 4 Distribution of the second level code data errors (type B error)

Code	*N*	%
2.1 Policy document link error	104	5.0
2.1.1 Link updated or expired	5	0.2
2.1.2 No page found	92	4.4
2.1.3 Link to page of multiple document	7	0.3
2.2 False positive policy document mention	261	12.6
2.2.1 Policy document mentioned by itself	57	2.7
2.2.2 Policy document has not mentioned the scientific product	204	9.8
2.3 Transcription error	1080	52.0
2.3.1 Title error of scientific product	71	3.4
2.3.2 Author error of scientific product	781	37.6
2.3.2.1 Omission	79	3.8
2.3.2.2 Misspell	21	1.0
2.3.2.3 Duplicate	681	32.8
2.3.3 Source (Journal) error of scientific product	45	2.2
2.3.4 Publication date error/inconsistency of scientific product	161	7.7
2.3.4.1 Omission of publication date	21	1.0
2.3.4.2 Inconsistency of publication date	139	6.7
2.3.5 Title error of policy document	22	1.1

N is Number of incorrect records

(Code 1.2) Title error of scientific product

The second relatively frequent type A error concerns the title of mentioned publication. Figure 4 (in the "Appendix") exemplifies that in the original publication the title is "Thrive and Survive" while in the policy document it becomes "Succeed".

(Code 1.3) Publication date error of scientific product

Publication date could also be wrongly recorded by the policy document. Figure 5 (in the "Appendix") shows that the mentioned publication is in fact published in 1999 but is written as 2009 in the policy document.

(Code 1.4) Source (Journal) error of scientific product

The least commonly seen type A error concern the source of mentioned publication. Figure 6 (in the "Appendix") shows that the journal *Progress in Development Studies* is put as *Progress of Development Studies* in the policy document.

Type B error

For type B error, as shown in Table 4, the *code 2.3 transcription error* is highlighted and takes up 52% of total erroneous records. It is the major reason of the astonishingly

high percentage of type B error. 13% of policy document mentions are false positive where the mentioned scientific product according to Altmetric database is in fact not mentioned by the policy document. The situation of policy document link error is also poor, 5% of the links are no longer accessible.

(Code 2.1.1) Policy document link is updated or expired

In a few cases the policy document link is updated or expired but the link is not updated in the Altmetric database. Figure 7 (in the "Appendix") exemplifies that some policy document is withdrawn and some policy document is updated to another location, as a result, the policy document cannot be visited via the link provided by the Altmetric database.

(Code 2.1.2) Policy document page is not accessible

The most frequent situation where policy documents are not available is just because the page cannot be found any more. Figure 8 (in the "Appendix") shows that the policy document page cannot be found from the source website.

(Code 2.1.3) Policy document page refers to multiple policy documents

In a few cases, the link from the Altmetric database leads to the page where multiple policy document can be found, but readers are unable to judge which one to look at, as shown in Fig. 9 (in the "Appendix").

(Code 2.2.1) Policy document has mentioned itself

This is a relatively severe kind of type B error. To our observation, it is usually because the policy document is also a scientific product and there is a suggested citation format for the policy document. The suggested citation is wrongly recognized as a policy document mention. Figure 10 (in the "Appendix") has exemplified the case. At the end of the policy document entitled with *Setting of new MRLs and import tolerances for fluopyram in various crops* there is its suggested citation and it is falsely counted as a policy document mention.

(Code 2.2.2) Policy document has not mentioned the scientific product

The relatively severe error is that the policy document in fact has not mentioned the scientific product at all. As an example, Fig. 11 (in the "Appendix") shows that according to Altmetric database, the policy document has mentioned the publication entitled with *Disability living allowance*, but this publication cannot be found in the original policy document, instead it merely happens to appear in the full-text. We also have observed other reasons. For example, Fig. 12 (in the "Appendix") shows that the DOI in the policy document is wrongly detected due to the line break and therefore is mistaken as another irrelevant publication. The DOI of mentioned publication is https://doi.org/10.1016/j.nrl.2016.08.006 but in Altmetric database it is recorded as 10.1016/j. Nevertheless, underlying reason for most of the false positive records are unknown. The supposed mentioned publication just cannot be found in the original policy document by all means.

(Code 2.3.1) Title error of the mentioned scientific product found in the altmetric database

Error can be found in title of the mentioned scientific product in the altmetric database. Figure 13 (in the "Appendix") exemplifies that part of the title of the publication is missing. The missing part is usually the part after colon, or in a few cases merely the latter part of a long title.

(Code 2.3.2) Author error of the mentioned scientific product found in the altmetric database

Author error of the mentioned scientific product is the predominant type B error. It is the major reason for the high percentage of type B error. Figure 14 (in the "Appendix") exemplifies the situation where some author of mentioned publication is omitted. The mentioned publication is co-authored by Podgursky M. and Tongrut R., but in Altmetric database Tongrut R. is omitted. Figure 15 (in the "Appendix") shows that the author Müller-Lissner is misspelled as Muller-Lissner. This is due to the failure of handling non-English letters. Figure 16 (in the "Appendix") shows that authors of the publication are listed twice in Altmetric database record. The authors are listed in full name once and listed in abbreviated name again.

(Code 2.3.3) Source (Journal) error of the mentioned scientific product found in the altmetric database

We see a few records that source title of the mentioned scientific product is inaccurately recorded in Altmetric database. Figure 17 (in the "Appendix") shows that the journal of mentioned publication is *Gender and Development* but in Altmetric database it is written as *Gender and Development.* This is probably due to improper encoding or decoding by the program.

(Code 2.3.4) Publication date error/inconsistency of the mentioned scientific product found in the altmetric database

Publication date errors are also found in Altmetric database. Figure 18 (in the "Appendix") shows that the mentioned publication was published in 2009 but the publication date is omitted in Altmetric database. Figure 19 (in the "Appendix") shows that the mentioned publication was published in 2013 but in Altmetric database the publication date is recorded as 2012.

Table 5 Comparison of type A errors introduced by authors of policy document and citing publications

Error by author of policy document		Error by author of citing publication (Buchanan 2006)	
Author error of mentioned scientific product	5.2%	Author error of cited publication	3.7%
Title error of mentioned scientific product	1.5%	Title error of cited publication	1.0%
Publication date error of mentioned scientific product	0.9%	Publication date error of cited publication	0.6%

(Code 2.3.5) Title error of the policy document found in the altmetric database

A relatively small percentage of type B error concerns the title of policy document. Figure 20 (in the "Appendix") shows that the title of policy document recorded in the Altmetric database is inconsistent with the original title.

Discussion and conclusion

Comparison of data accuracy between policy document mention and citation data

The data quality of policy document altmetrics is compared with that of citation data. Authors of citing publication may produce errors to data of cited publication, similarly, authors of policy document may produce errors to data of mentioned scientific product. Citation database uses program to automate the identification and extraction of citation data, similarly, altmetrics database uses program to crawl and collect data of policy document, but altmetrics database usually gets bibliographic data directly from the third-party citation database instead of collecting the bibliographic data on their own. In case of Altmetric database, CrossRef is the major bibliographic data provider.

For type A error, Table 5 has compared the error rate caused by author of policy document and author of citing publication. It is found that percentage of data errors of mentioned scientific product in policy document are all slightly higher than that of cited publications in citing publications, indicating no significant difference.

For type B error, because sub-categories of error in altmetrics database are different from that in citation database, they are not directly comparable. However, the transcription error (52%) of altmetrics database is obviously much higher than that of citation database which is 3.5% according to Franceschini et al. (2016b).

Possible reasons and influence of the data errors

To begin with, errors in this study are defined from user's perspective. Suppose an end user of altmetric database is browsing the database, the user would probably be unable to figure out the underlying reasons of the inconsistencies and regard them as errors. There could be other perspectives, for example, from researcher's perspective, the first online date as the publication date is sometimes more accurate in reflecting the velocity of data accumulation (Fang and Costas 2020). Or, from altmetric database's perspective, bibliographic data are correct as long as they are in consistency with the data provided from the third party,

because it is not their duty to check the accuracy of the bibliographic data. Based on the definition of error from user's perspective, we further discuss the possible reasons and their influences.

Type A errors are from the authors of the policy documents. Therefore, the underlying reason for type A error is the incaution of authors of policy document. Omission and misspell of title, author or publication date in the references could be attributable to this reason. Interestingly, sometimes altmetrics databases are able to eliminate such kind of error, because Altmetric database doesn't extract bibliographic data of the mentioned publication from the policy document, instead they use third-party citation database which extract bibliographic data from the source publications. For this reason, in practical usage of altmetrics data, type A error seems to be not a big problem,

Type B errors are further classified into three categories, error code 2.1 (policy document link error) is related to the provenance of policy document, error code 2.2 (false positive policy document mention) is related to the false positive mentions, error code 2.3 (transcription error) is related to the metadata quality of the mentioned records.

For error code 2.1, the underlying reason may be the lack of updating the old records. Due to the dynamic nature of policy document websites, the policy document could be revised and thus updated to a new URL. In other cases, the policy document could be retracted or the entire website section is no longer active. Policy document link error (percentage is 5%) has corroborated that altmetrics data are highly influenced by the stability of source web platform. Once the platform has removed the link or retracted the policy document, the policy altmetrics data are gone forever. Theoretically, Altmetric database could update periodically and revisit the policy document link to check whether it is still available or linking to the right policy document, but it may not be economically feasible for the company. Therefore, this type of error is difficult to get rid of.

Moreover, according to the comments from experts of Altmetric.com company, Altmetric database has chosen to keep records of which the policy document link is no longer accessible, because policymakers sometimes remove policy documents for very good reasons (e.g. clinical care guidelines issued by UK's NICE that go out of date). Those citations are kept in the database because researchers typically want to know if they'd been cited previously, even if those documents are no longer available. In a way, it's akin to citation databases indexing citations from print journals that aren't available online. There's certainly an element of trust that the citation found previously was correct. As a result, error code 2.1 would not be a severe problem but will cause inconvenience if content analysis of policy document is to be conducted.

For error code 2.2, the underlying reason may be bugs in the program for processing the data. Some of the bugs are minor and easy to get fixed, for example bugs that are related to encoding or decoding. These bugs make the special letter like "&" not properly presented. Some of the bugs are severe and could be easily fixed, for example, the omission or duplication of authors, the error is probably related to the underlying data merging process. Meanwhile, some of the bugs are severe and not easy to find out, for example, the DOI could be wrongly extracted from the policy document file when there is line break in the DOI string. The program is also unaware of the fact that some policy document gives its suggested citation at the end of the text and wrongly identifies that as a policy document mention. It can be avoided by excluding self-mentions. Error code 2.2 could bring very misleading results, because many related altmetrics analysis are based on the count of policy document mentions. The percentage is 12.6% and cannot be neglected.

For error code 2.3, the underlying reason may be attributable to the third-party bibliographic data providers. Therefore, Altmetric database could report these errors to the

bibliographic data providers and help them improve the metadata quality. In error code 2.3, publication date error could be difficult to define, because a scientific product could have multiple publication date (Haustein et al. 2015). However, in this study, publication date in the published journal is used as standard publication date, because it is commonly used for selecting data for scientometric analysis. Consequently, many records in altmetrics database that use first online mention date as publication date are deemed erroneous, although they may be more accurately regarded as inconsistency. The error code 2.3 is trivial or can even be neglected for conducting altmetrics analysis, because researchers commonly use bibliographic data from traditional citation database, only unique identifiers (e.g. DOI) provided by Altmetric database are used to match for corresponding altmetrics data. But error code 2.3 could cause potential inaccuracy if metadata are used to retrieve in the Altmetric database and later used directly for analysis.

Further remarks

Results of the study have made it clear that altmetrics data, especially policy document altmetrics data, are far from being perfect. Future studies based on the data need to be aware of this limitation. Although the overall error rate in Altmetric database is much higher than that of traditional citation database, majority of the errors are produced either by the policy document website or the third-party bibliographic data provider. These errors are trivial and can even be neglected in conducting altmetrics research. Moreover, in June 2019 we did a random but not systematic re-check of the erroneous records, and found that most errors related to bibliographic information, such as omission of publication date, duplication of authors and misspell of journal title, seem to be gone.

The error code 2.2 (false positive policy document mention) is perhaps the only error type that could be directly improved by the Altmetric database. The error type would have negative influence on altmetrics research. In June 2019, the random check showed the false positive policy document mentions were still there. For code 2.2.1 error, it can be avoided by removing self-citing policy document. For code 2.2.2 error, more studies are needed to find out the underlying reasons. However, considering the high complexity of extracting mentions of publications from various sources and formats of policy documents, Altmetric database has achieved quite excellent performance.

Due to high labor cost of manual coding work, the study is based on analysis of a relatively small sample of the policy document altmetrics data. This also needs to be taken into consideration in interpreting the results of the study.

Acknowledgements The present study is an extended version of an article (Yu et al. 2019) presented at the 17th International Conference on Scientometrics and Informetrics, Rome (Italy), 2–5 September, 2019. The authors would like to thank Longfei Li and Zihan Yin for helping conduct the coding work. The authors would like to thank Altmetric.com company for providing access to the data and Stacy Konkiel for the useful comments. The research is supported by National Natural Science Foundation of China (NO.71804067), Humanity and Social Science Foundation of Ministry of Education of China (18YJC870023), the Fundamental Research Funds for the Central Universities (No. 30920021203).

Appendix

See Tables 6 and 7; Figs. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 and 20.

Table 6 Coding table for type A error (errors found in the policy document)

Code	Definition
1.1 Author error of scientific product	Author of scientific product wrongly recorded in the policy document
1.2 Title error of scientific product	Title of scientific product wrongly recorded in the policy document
1.3 Publication date error of scientific product	Publication date of scientific product wrongly recorded in the policy document
1.4 Source (Journal) error of scientific product	Source (Journal) of scientific product wrongly recorded in the policy document

Table 7 Coding table for type B error (errors found in altmetrics database)

Code	Definition
2.1 Policy document link error	
2.1.1 Link updated or expired	Link to the policy document is updated or expired
2.1.2 No page found	Link to the policy document returns no page found
2.1.3 Link to page of multiple document	Link to the policy document leads to multiple versions of files
2.2 False positive policy document mention	
2.2.1 Policy document mentioned by itself	Mentioned scientific product is the recommended citing format of the policy document itself
2.2.2 Policy document has not mentioned the scientific product	The scientific product cannot be found in the source policy document
2.3 Transcription error	
2.3.1 Title error of scientific product	Title error of scientific product in the database
2.3.2 Author error of scientific product	Author error of scientific product in the database
2.3.2.1 Omission	Author's name omitted
2.3.2.2 Misspell	Author's name misspelled
2.3.2.3 Duplicate	Author's name duplicated
2.3.3 Source (Journal) error of scientific product	Source (Journal) title error of scientific product in the database
2.3.4 Publication date error/inconsistency of scientific product	
2.3.4.1 Omission of publication date	Publication date omitted in the database
2.3.4.2 Inconsistency of publication date	Publication date inconsistent with the formal publication date
2.3.5 Title error of policy document	Title error of policy document in the database

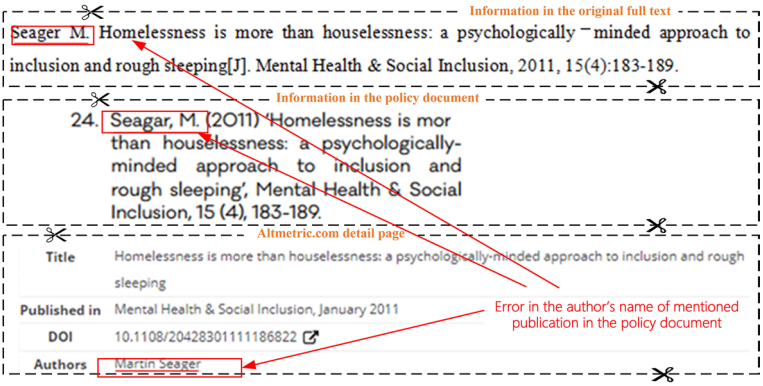


Fig. 3 Example of type A error due to the inaccurate author names of mentioned publication. This error is classified to sub-category 1.1. The Altmetric database was queried in June 2018

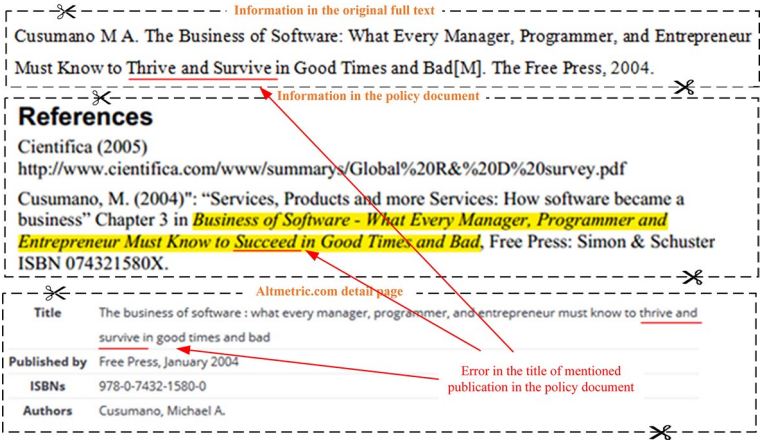


Fig. 4 Example of type A error due to the inaccurate title of mentioned publication. This error is classified to sub-category 1.2. The Altmetric database was queried in June 2018

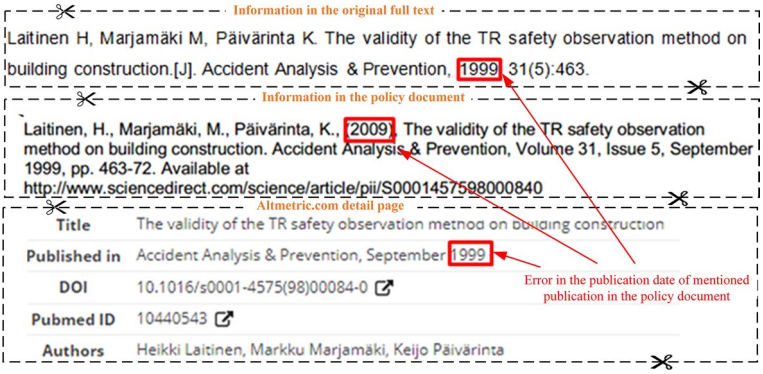


Fig. 5 Example of type A error due to the inaccurate publication date of mentioned publication. This error is classified to sub-category 1.3. The Altmetric database was queried in June 2018

Fig. 6 Example of type A error due to the inaccurate title of source journal. This error is classified to sub-category 1.4. The Altmetric database was queried in June 2018

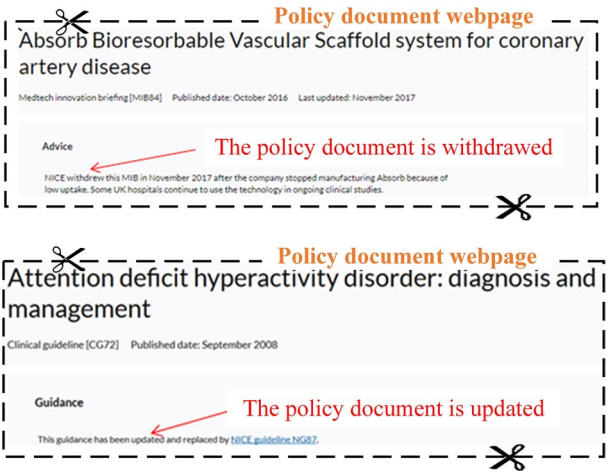


Fig. 7 Example of type A error due to that policy document link is updated or expired. This error is classified to sub-category 2.1.1. The Altmetric database was queried in June 2018

Fig. 8 Example of type B error due to that the policy document page cannot be viewed. This error is classified to sub-category 2.1.2. The Altmetric database was queried in June 2018

Fig. 9 Example of type B error due to that multiple policy document are provided in the page. This error is classified to sub-category 2.1.3. The Altmetric database was queried in June 2018

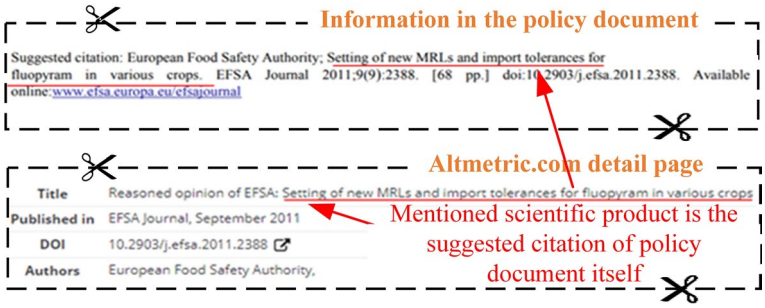


Fig. 10 Example of type B error due to that policy document has mentioned itself. This error is classified to sub-category 2.2.1. The Altmetric database was queried in June 2018

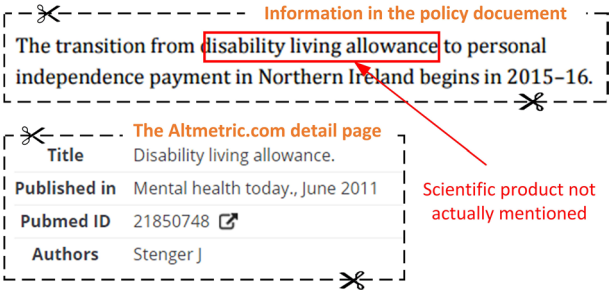


Fig. 11 The first example of type B error that policy document has not mentioned the scientific product. This error is classified to sub-category 2.2.2. The Altmetric database was queried in June 2018

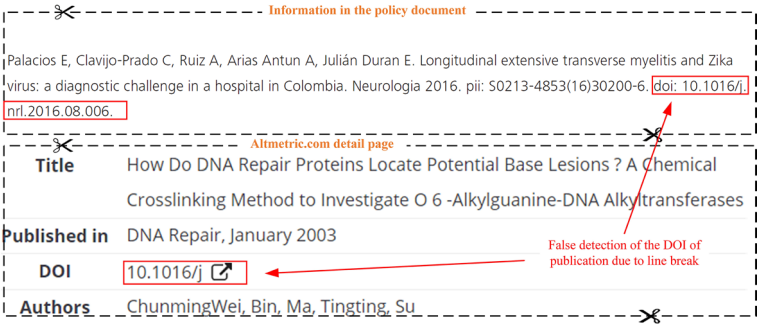


Fig. 12 The second example of type B error that policy document has not mentioned the scientific product. This error is classified to sub-category 2.2.2. The Altmetric database was queried in June 2018

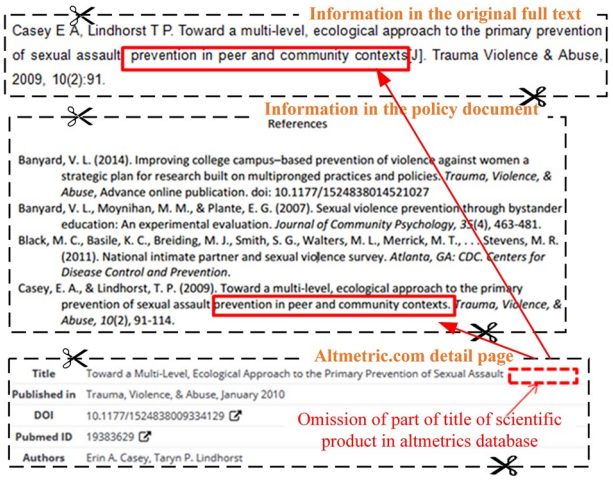


Fig. 13 Example of type B error that title of mentioned scientific product is incorrect in the Altmetric database. This error is classified to sub-category 2.3.1. The Altmetric database was queried in June 2018

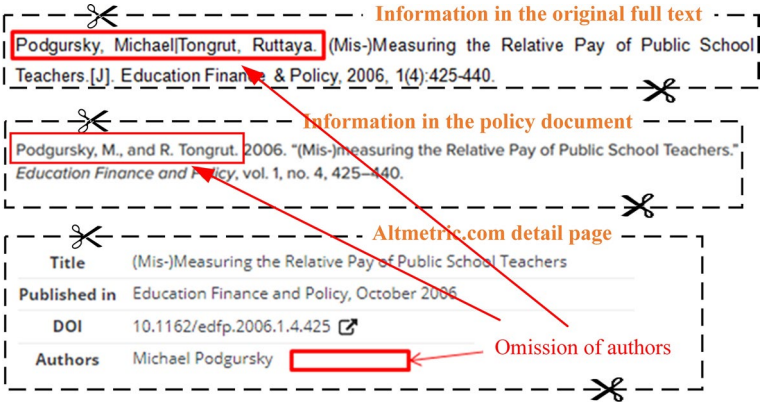


Fig. 14 Example of type B error that author of mentioned scientific product is omitted. This error is classified to sub-category 2.3.2.1. The Altmetric database was queried in June 2018

Fig. 15 Example of type B error that authors of mentioned scientific product are misspelled. This error is classified to sub-category 2.3.2.2. The Altmetric database was queried in June 2018

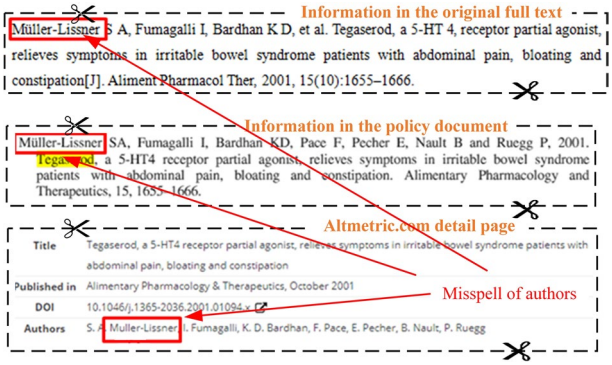
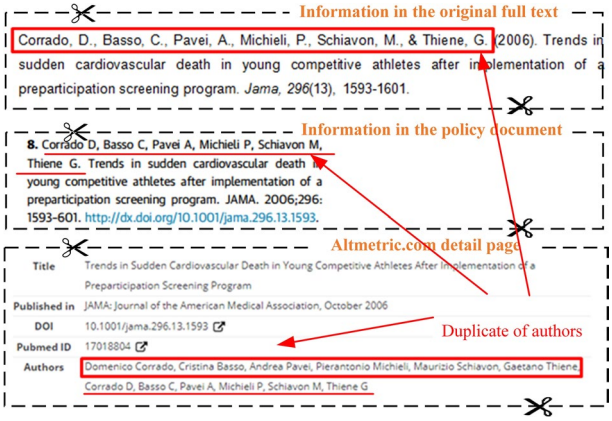


Fig. 16 Example of type B error that author information of mentioned scientific product is duplicated. This error is classified to sub-category 2.3.2.3. The Altmetric database was queried in June 2018



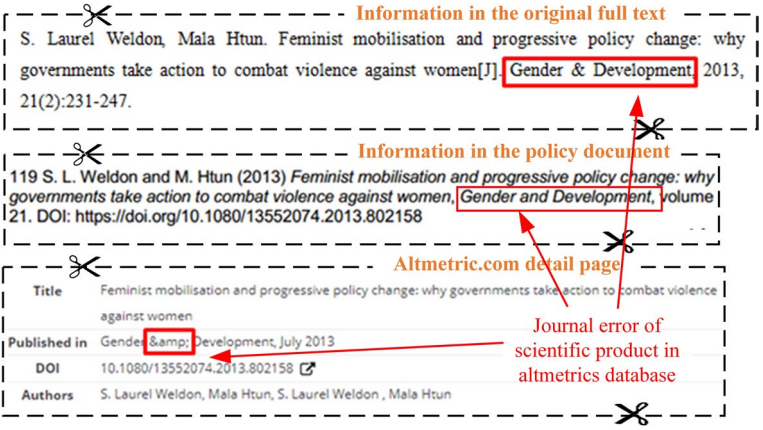


Fig. 17 Example of type B error that journal title of scientific product is incorrect in altmetrics database. This error is classified to sub-category 2.3.3. The Altmetric database was queried in June 2018

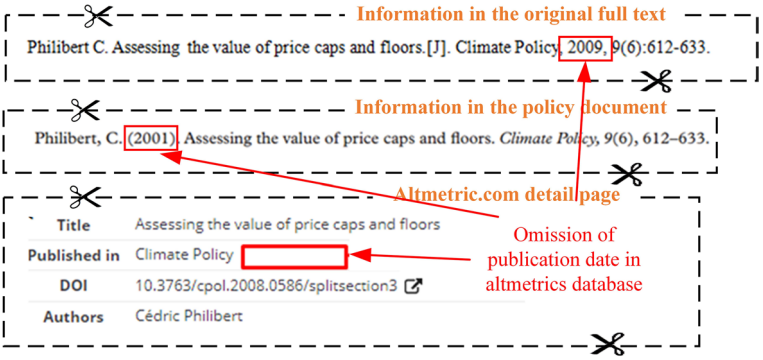


Fig. 18 Example of type B error that publication date of mentioned publication is omitted. This error is classified to sub-category 2.3.4.1. The Altmetric database was queried in June 2018

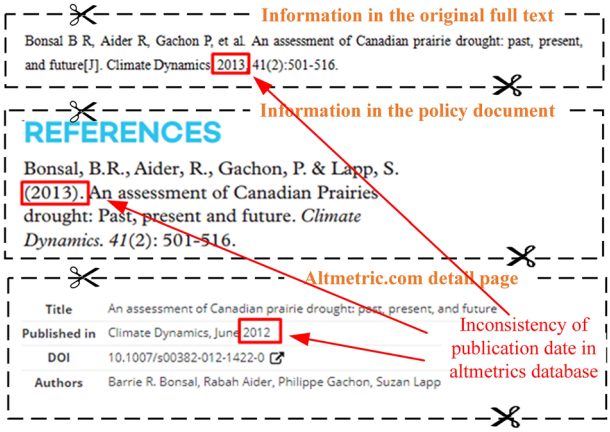


Fig. 19 Example of type B error that publication date of mentioned publication recorded in Altmetric database is inconsistent with the original publication. This error is classified to sub-category 2.3.4.2. The Altmetric database was queried in June 2018

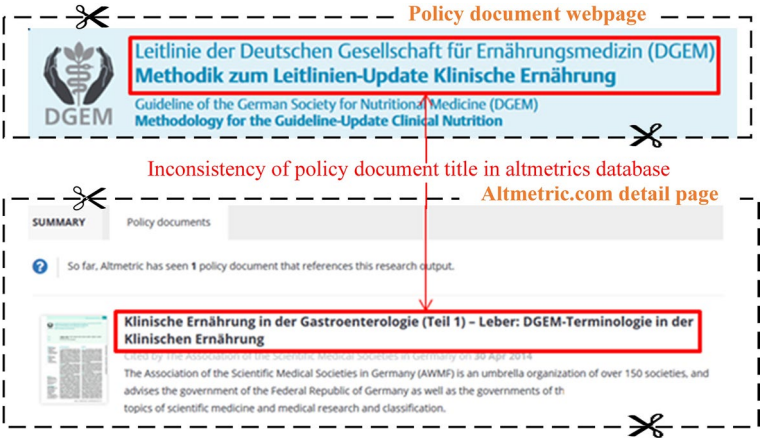


Fig. 20 Example of type B error that title of policy document recorded in Altmetric database is inconsistent with the source policy document. This error is classified to sub-category 2.3.5. The Altmetric database was queried in June 2018

References

Aduku, K. J., Thelwall, M., & Kousha, K. (2017). Do Mendeley reader counts reflect the scholarly impact of conference papers? An investigation of computer science and engineering. *Scientometrics, 112*(1), 573–581.

Archambault, É., Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the American Society for Information Science and Technology, 60*(7), 1320–1326.

Bar-Ilan, J. (2014). JASIST@ Mendeley revisited. altmetrics14: Expanding impacts and metrics. In *Workshop at web science conference 2014*.

Bornmann, L., & Haunschild, R. (2015). Which people use which scientific papers? An evaluation of data from F1000 and Mendeley. *Journal of Informetrics, 9*(3), 477–487.

Bornmann, L., Haunschild, R., & Marx, W. (2016). Policy documents as sources for measuring societal impact: How often is climate change research mentioned in policy-related documents? *Scientometrics, 109*(3), 1477–1495.

Buchanan, R. A. (2006). Accuracy of cited references: The role of citation databases. *College and Research Libraries, 67*(4), 292–303.

Calver, M. C., Goldman, B., Hutchings, P. A., & Kingsford, R. T. (2017). Why discrepancies in searching the conservation biology literature matter. *Biological Conservation, 213,* 19–26.

Chamberlain, S. (2013). Consuming article-level metrics: Observations and lessons. *Information Standards Quarterly, 25*(2), 4–13.

Chen, X. J. (2014). Study of quality of references and its auditing methods. *Chinese Journal of Scientific and Technical Periodicals, 25*(9), 1145–1148. **(in Chinese)**.

Donner, P. (2017). Document type assignment accuracy in the journal citation index data of Web of Science. *Scientometrics, 113*(1), 219–236.

Fang, Z., & Costas, R. (2020). Studying the accumulation velocity of altmetric data tracked by Altmetric.com. *Scientometrics, 123,* 1077–1101. https://doi.org/10.1007/s11192-020-03405-9.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2014). Scientific journal publishers and omitted citations in bibliometric databases: Any relationship? *Journal of Informetrics, 8*(3), 751–765.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016a). The museum of errors/horrors in Scopus. *Journal of Informetrics, 10*(1), 174–182.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016b). Empirical analysis and classification of database errors in Scopus and Web of Science. *Journal of Informetrics, 10*(4), 933–953.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016c). Do Scopus and WoS correct old omitted citations? *Scientometrics, 107*(2), 321–335.

Garfield, E. (1974). Errors-theirs, ours and yours. In *Essays of an information scientist* (Philadelphia: ISI Pr., 1977), 2: 80–81. Originally published in Current Contents (June 19, 1974): 5–6.
Harzing, A.-W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics, 106*(2), 787–804.
Haunschild, R., & Bornmann, L. (2017). How many scientific papers are mentioned in policy-related documents? An empirical investigation using Web of Science and Altmetric data. *Scientometrics, 110*(3), 1209–1216.
Haustein, S. (2016). Grand challenges in altmetrics: Heterogeneity, data quality and dependencies. *Scientometrics, 108*(1), 413–423.
Haustein, S., Bowman, T. D. & Costas, R. (2015). *When is an article actually published? An analysis of online availability, publication, and indexation dates*. Retrieved from https://arxiv.org/ftp/arxiv/papers/1505/1505.00796.pdf.
Khazragui, H., & Hudson, J. (2015). Measuring the benefits of university research: Impact and the REF in the UK. *Research Evaluation, 24*(1), 51–62.
Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology, 58*(13), 2105–2125.
Meschede, C., & Siebenlist, T. (2018). Cross-metric compatability and inconsistencies of altmetrics. *Scientometrics, 115*(1), 283–297.
Moed, H. F. (2002). The impact-factors debate: The ISI's uses and limits. *Nature, 415*(6873), 731.
Mohammadi, E., Kwasny, M., & Holmes, K. L. (2018). Academic information on Twitter: A user survey. *PLoS ONE, 13*(5), e0197265.
Newson, R., King, L., Rychetnik, L., et al. (2018). Looking both ways: A review of methods for assessing research impacts on policy and the policy utilisation of research. *Health Research Policy and Systems, 16*(1), 54.
NISO. (2016). *Outputs of the NISO alternative assessment project*. Retrieved from https://groups.niso.org/apps/group_public/download.php/17091/NISO+RP-25-2016+Outputs+of+the+NISO+Alternative+Assessment+Project.pdf.
Ortega, J. L. (2018). Reliability and accuracy of altmetric providers: A comparison among Altmetric.com, PlumX and Crossref Event Data. *Scientometrics, 116*(3), 2123–2138.
Ortega, J. L. (2019). Blogs and news sources coverage in altmetrics data providers: A comparative analysis by country, language, and subject. *Scientometrics*. https://doi.org/10.1007/s11192-019-03299-2.
Peters, I., Jobmann, A., Eppelin, A., et al. (2014). Altmetrics for large, multidisciplinary research groups: A case study of the Leibniz Association. *Libraries in the Digital Age (LIDA) Proceedings, 13,* 1–9.
Prins, A. A., Costas, R., van Leeuwen, T., & Wouters, P. F. (2016). Using Google Scholar in research evaluation of humanities and social science programs: A comparison with Web of Science data. *Research Evaluation, 25*(3), 264–270.
Su, X. N. (2001). Quality control of data in citation indexes. *Journal of Library Science in China, 27*(2), 76–78. **(in Chinese)**.
Tattersall, A., & Carroll, C. (2018). What can altmetric.com tell us about policy citations of research? An analysis of Altmetric.com data for research articles from the University of Sheffield. *Frontiers in Research Metrics and Analytics*, *2*(9), 1–11.
Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics, 10*(2), 347–364.
Wooldridge, J., & King, M. B. (2018). Altmetric scores: An early indicator of research impact. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24122.
Yu, H. (2017). Context of altmetrics data matters: An investigation of count type and user category. *Scientometrics, 111*(1), 267–283.
Yu, H., Cao, X., Xiao, T., & Yang, Z. (2019). Accuracy of policy document mentions: The role of altmetrics databases. In *Proceedings of ISSI 2019—The 17th international conference on scientometrics and informetrics* (pp. 477–488). Italy: Sapienza University.
Zahedi, Z., & Costas, R. (2018). General discussion of data quality challenges in social media metrics: Extensive comparison of four major altmetric data aggregators. *PLoS ONE, 13*(5), e0197326.
Zahedi, Z., Fenner, M., & Costas, R. (2014). How consistent are altmetrics providers? Study of 1000 PLOS ONE publications using the PLOS ALM, Mendeley and Altmetric.com APIs. altmetrics 14. In *Workshop at the Web Science conference, Bloomington, USA*.
Zahedi, Z., & Haustein, S. (2018). On the relationships between bibliographic characteristics of scientific documents and citation and Mendeley readership counts: A large-scale analysis of Web of Science publications. *Journal of Informetrics, 12*(1), 191–202.

Zahedi, Z., Haustein, S., & Bowman, T. (2014). Exploring data quality and retrieval strategies for Mendeley reader counts. In *SIG/MET workshop, ASIS&T 2014 annual meeting, Seattle*. Retrieved from: www.asis.org/SIG/SIGMET/data/uploads/sigmet2014/zahedi.pdf.

Zhao, Q. M. (2009). Analysis of errors in refereces of scientific journals and the prevention strategies. *Editorial Friends, 6,* 47–49. **(in Chinese)**.