The indexation of retracted literature in seven principal scholarly databases: A coverage comparison of Dimensions, OpenAlex, PubMed, Scilit, Scopus, The Lens and Web of Science

José Luis Ortega[1] (ORCID: 0000-0001-9857-1511) and Lorena Delgado-Quirós (ORCID: 0000-0001-8738-7276)

Institute for Advanced Social Studies (IESA-CSIC), Córdoba, Spain

Joint Research Unit Knowledge Transfer and Innovation, (UCO-CSIC), Córdoba, Spain

jortega@iesa.csic.es ; ldelgado@iesa.csic.es

Abstract

In this study, the coverage and overlap of retracted publications, retraction notices and withdrawals are compared across seven significant scholarly databases, with the aim to check for discrepancies, pinpoint the causes of those discrepancies, and choose the best product to produce the most accurate picture of retracted literature. Seven scholarly databases were searched to obtain all the retracted publications, retraction notices and withdrawal from 2000. Only web search interfaces were used, excepting in OpenAlex and Scilit. The findings demonstrate that non-selective databases (Dimensions, OpenAlex, Scilit, and The Lens) index a greater amount of retracted literature than do databases that rely their indexation on venue selection (PubMed, Scopus, and WoS). The key factors explaining these discrepancies are the indexation of withdrawals and proceeding articles. Additionally, the high coverage of OpenAlex and Scilit could be explained by the inaccurate labeling of retracted documents in Scopus, Dimensions, and The Lens. 99% of the sample is jointly covered by OpenAlex, Scilit and WoS. The study suggests that research on retracted literature would require querying more than one source and that it should be advisable to accurately identify and label this literature in academic databases.

Keywords: retracted publications, retraction notices, withdrawals, scholarly databases, coverage, overlap

1. Introduction

With the emergence of new journal types (i.e., open access, paywall, hybrid journals), commercial strategies (i.e., subscription, APCs, diamond), and peer review models (i.e., open and blind review), the current publishing system is undergoing a significant transformation that is raising questions about how these changes may affect research integrity (Barber, 2021; Sanderson, 2023). These concerns have increased interest in studying the correction of science using bibliometric techniques. Analyzing the frequency of retractions and retracted papers in scientific literature is one such example.

[1] Corresponding author: Institute for Advanced Social Studies (IESA-CSIC), Camposanto de los Mártires, 7 14004 Córdoba, Spain jortega@iesa.csic.es

Although not all retractions may be the result of instances of misconduct, their number has been used as a measure for the correction of science.

However, there are significant methodological issues with the analysis of this type of publications. The main factor is that the publication of retraction notices and their connection to the original publication are not normalized. Despite the recommendations made by COPE (Committee on Publication Ethics) in 2019, each publisher, even each journal, chooses how to notify readers when a publication is withdrawn. In the best cases, journals publish a retraction notice explaining the retraction's circumstances and labeling the retracted document to alert readers. But in many other instances, the original publication is replaced by the retraction notice, just marked as retracted, or even deleted for no apparent reason (Vuong, 2020). Sometimes, retraction notices are concealed within a journal issue without a specific identifier or inserted after another article. These procedures make it difficult to find retractions notices and determine when and why articles are withdrawn. This circumstance complicates tracking this kind of literature accurately and distorts estimates of erroneous or dishonest research.

All of these issues provide a significant barrier for bibliographic databases that cover scientific literature. These tools must not only correctly index retractions but also update their records when an article is retracted and link the retraction notice to the article in question. Many databases today show linkages between retracted publications and retraction notices, although not all records show these relationships, and these connections are not always clear from the metadata. The inability to recognize and connect publications influences their coverage, which in turn affects the perception about the incidence of retracted publications.

These problems are of special interest to their users, because they need to correctly and promptly determine when an article has been notified or removed, including the reasons of this problem. This information is essential to prevent the citation of erroneous literature in reviews and meta-analyses (Hsiao & Schneider, 2021; Bolland et al., 2022), as well as to impede the validation and diffusion of unfound theories (Suelzer et al., 2019).

The establishment of Crossref as an open repository for publisher metadata, the unrestricted release of search engine data (Microsoft Academic Graph, CiteSeerX, AMiner), as well as other open initiatives (PubMed, DOAJ), have led to the emergence of new third-party academic databases (Dimensions, Scilit, The Lens, OpenAlex), which increase the visibility of these problematic publications and offer new options for handling them. The advent of so many alternative products (e.g., Crossref-Retraction Watch, COCI) increases the need to understand what documents are covered, how the data is managed, and any potential retrieval issues. This study attempts to address these questions, by exploring how this information is retrieved in each database and what variations in coverage there are between them.

2. Literature review

Literature around the correction of science has been extensive and diverse because there are important limitations to understanding how misconduct might be detected in publications. For instance, the use of retractions as a proxy for misconduct have led to different or incomplete interpretations. According to Steen (2011), the consistent increase in retractions indicated that misconduct levels appeared to be higher than in the past. Fanelli (2013), however, stated that increased editorial board scrutiny was mostly the cause of that rise. Furthermore, some studies have found that there should be a far higher percentage of misconduct in publications than what is disclosed in editorial notices (Cokol et al., 2007; Stricker and Günther, 2019). Bik et al. (2018) manually examined image manipulation in molecular and cellular biology journals, and they discovered that only 10% of the studies that showed evidence of image manipulation were retracted. Ortega and Delgado-Quirós (2023) pointed out that only 21.5% of publications on PubPeer that were accused of misconduct received an editorial notice. Otherwise, a sizable percentage of retractions are not the result of misconduct or fraud. The first study to examine the content of retractions was conducted by Budd et al. (1998), who discovered that 37% of retractions were the result of clear wrongdoing. Decullier et al. (2013) found that the two most common reasons for retraction were fraud (14%), and plagiarism (20%). Most recently, Lei and Zhang (2018) found that three out of every four retractions in China are the result of misconduct. All these studies demonstrate how difficult it is to precise and identify scientific misconduct cases using retractions.

The selection of the database to be utilized for analyzing retractions must be given careful thought. The first studies about the incidence of retractions were already published at the turn of the 20th Century and were focused on biomedical literature (Snodgrass & Pfeifer, 1992; Budd et al., 1998). Medline, or its web interface PubMed, were the primary sources used to estimate the rise in retracted literature (Nath et al., 2006; Redman et al., 2008). Up to 2013, PubMed was the only database for learning about retractions, with the exception of Trikalinos et al. (2008) who used the Web of Science (WoS) to describe the retraction difficulties in high-impact journals (Hesselmann et al., 2017). The main reason is that, up until that point, PubMed was the most trustworthy source for labeling and connecting retractions and retracted publications. This circumstance demonstrates how crucial it is for scientific databases to accurately identify retractions to perform this kind of studies. Thus, for instance, the number of studies using WoS began to raise from 2012 (Grieneisen & Zhang, 2012; Fanelli, 2013; Lu et al., 2013), when this database likely started to identify these documents. Later, Scopus database also commenced to be utilized for retrieving retracted publications (Aspura et al., 2018; Elango et al., 2019). Today, the recent proliferation of bibliographic databases has fueled extensive studies that offer a more in-depth analysis about the background of erroneous or fraudulent science. Ribeiro and Vasconcelos (2018) were the first to measure the prevalence of retractions by country using Retraction Watch. Using open citation indexes (Microsoft Academic and COCI), Heibi and Peroni (2022) monitored the number of citations received by retracted humanities papers. They were able to determine from these sources that retracted works in the humanities did not have a decrease in citations after retraction. Crossref, Dimensions, and Netscity were utilized by Cabanac et al. (2023) to map the locations of the cities with the highest percentage of retractions. And Malkov et al. (2023) studied the spread of retractions within policy documents using altmetric providers such as Altmetric.com and Overton.

Several studies comparing the coverage of retractions and retracted papers have emerged along with the quick development of various scholarly databases. Schmidt (2018) was the first one to examine how PubMed and WoS label retracted publications. She discovered that a one third of PubMed retracted publications and retractions were not labeled in WoS. The coverage of retracted Korean publications in Scopus and KoreanMed was examined by Kim et al. (2019). Their findings demonstrated that Scopus had indexed all of the KoreanMed records. Proescholdt and Schneider (2020) investigated how retracted documents are correctly identified in Pubmed, WoS, and Scopus. They discovered that many retracted publications in Scopus were not correctly labeled. When Uppala et al. (2022) evaluated the editorial notice coverage across Crossref, PubMed, and Scite, they found significant variations in how retracted publications and retraction notices were classified. The broadest comparison analysis to date was carried out by Schneider et al. (2023) using Crossref, Retraction Watch, Scopus, and Web of Science. Only 3% of shared papers are classified as retractions across all sources, according to their findings.

However, there is not a yet a clear knowledge how recent scholarly databases, with free-access interfaces (i.e. Dimensions, OpenAlex, PubMed, Scilit, The Lens), perform in comparison to classical sources such as WoS and Scopus. In which form they cover retracted publications and retraction notices, how these publications are labeled and updated, and what advantages or disadvantages there are to retrieving these records with the highest recall.

3. Objective

This study attempts to analyze the coverage of retracted publications, retraction notices, and withdrawals in seven bibliographic databases (Dimensions, OpenAlex, PubMed, Scilit, Scopus, The Lens and WoS), with the aim of examining, firstly, how these publications are identified and labeled, and as this will affect how they are retrieved; and, secondly, determine which source or combination of sources has the broadest coverage. The following research questions were developed:

- What coverage variations are there between databases?
- What variables could account for specific coverage and overlap variations between databases?
- Which database combination gives the most accurate image of retracted articles, retraction notices and withdrawals?

4. Methods

To compare the coverage of retractions and retracted or withdrawn publications we have developed a quantitative methodology based firstly on search in a varied range of bibliographic databases for this type of publications. Then, the entire list of retrieved documents is compared again with these databases to test to what extent they index the complete pull of publications.

4.1. Definitions

First, we must define the precise meaning of the examined document types to identify their particular characteristics and how they could affect search and retrieval. Our definitions are based on COPE guidelines (2019).

Retracted publication: This is the original retracted publication. This publication should include a label that reads "Retracted" after it has been retracted, along with a link to the retraction notice. In many cases, the retracted publication is replaced by the retraction notice, which might be challenging to determine when a paper was retracted.

Retraction notice: it is an independent publication that alert readers when one or more articles have been retracted. The cause of the retraction should be stated in this notification, together with a reference to and identification of the original retracted publication. Retraction notices are frequently used to update retracted publications; as a result, both the retraction notice and the retracted article have the same identifier (e.g., doi).

Withdrawal: Publishers occasionally refer to publications as being "withdrawn" when they have errors or have behaved improperly. Elsevier only allows publications to be withdrawn when they are not still assigned to an issue and are online. However, many other publishers define this action as a retraction, removal or a publishing mistake. COPE guidelines (2019) do not mention this correction type. In most cases, the withdrawn implies the removal of the original publication by a brief note.

4.2. Sources

Seven bibliographic databases were chosen to retrieve bibliographic information about retractions, and withdrawn and retracted publications. These sources were selected because they provide a search interface for retrieving these specific types of documents and they are not specialized in this type of publications. Specialized products in retractions such as Retraction Watch or OpenRetractions were excluded from the study because we are unable to evaluate how they label retraction notices, retracted publications and withdrawals nor how their search interfaces can filter these publications. That would produce a selection bias in favor of the specialized products.

Dimensions (app.dimensions.ai): This search service was developed by Digital Science in 2018 and cover around 130 million of scientific papers. It is based on outside sources, mainly CrossRef and PubMed. Despite the fact that Dimensions lacks a document type for recognizing retractions and retracted or withdrawn publications, it occasionally contains a notice when a paper is retracted that includes a link to the publisher's website.

OpenAlex (openalex.org): This is the newest product; it debuted in 2022. It is a nonprofit endeavor with the goal of developing an accessible bibliographic database for the academic community. The Microsoft Academic Graph, an open release of the former Microsoft Academic, serves as the foundation of OpenAlex and has been enhanced with information from additional open sources including Crossref and PubMed. It indexes about 240 million publications in total. To detect papers that have been retracted, OpenAlex has a binary field (*is_retracted:true*). Although this information is verified by Crossref, searches by title were also carried out.

PubMed (pubmed.ncbi.nlm.nih.gov): This is the only bibliographic database with a focus on the field of biomedicine. Developed by the National Library of Medicine in 1996, it is a web interface to search MEDLINE database, which contains more than 34 million references. Not only does this service distinguish between retractions and retractable articles, but it also links these documents together. For the retrieval of retraction notices and retracted articles, it employs the labels *retraction of publication* and *retracted publication*.

Scilit (www.scilit.net): This database was created by MDPI publisher in 2014. It covers 149 million of publications from Crossref, PubMed, preprint repositories and publishers. Scilit labels and links retracted publications with their retraction notices (*retraction*, *withdrawal*), as well as to notice when a publication has been withdrawn.

Scopus (www.scopus.com): It is one of the most important citation indexes, which was developed by Elsevier in 2004. Scopus builds a database of 78 million records by obtaining data directly from the publishers. Scopus occasionally associates retractions with retracted publications. It has a document type for retracted publications (*tb*), but it lacks a document type for withdrawals and retractions. In that instance, the publication's title was also searched for the word *retraction* and *withdrawn*.

The Lens (www.lens.org): Lens is a service for discovering patents and scholarly publications produced by Cambia in 2013. As Dimensions, this database adds entries from Crossref, PubMed, Core and OpenAlex, which leads to collect more than 247 million of bibliographic records. The Lens only distinguishes between these types of publications by the labels (such as *retracted*, *withdrawn*, and *retraction notice*) that are included in the document's title.

Web of Science (WoS) (webofscience.com): It is the web platform for the different citation indexes created by the Institute for Information Science (now Clarivate) in 1964. Launched in 1997, Web of Science has 193 million of documents, 85 of which are in its core collection. This portal assigns *retraction* category to retraction notices and *retracted publication* to retracted documents. Moreover, it adds the label *Retracted* to identify publications that have been retracted. It now also connects retracted documents to the associated retraction notice.

4.3. *Data collection*

The data retrieval and search process were finished in June 2022. Because the number of retractions and retracted documents is less frequent and their link is difficult to track, all queries were restricted to only retrieve information from 2000 onward. We searched and downloaded the material via the online search interface because this endpoint offers a reliable syntax for retrieving this type of publications. Only in the cases of OpenAlex and Scilit we have made use of a REST API endpoint, either because the application did not offer a web interface (OpenAlex) or because the query syntax was identical to that of the web interface (Scilit). The data extraction procedure is summarized in Table 1 with the endpoints, queries, and results broken down each database.

Database	Access point	Queries	Result	After cleansing

Dimensions	Simple search (https://app.dimensions.ai/discover/publication)	retracted OR retraction OR withdraw* FILTERS > Publication year>=2000	53,644	47,913 (89.3%)
The Lens	Structured search (https://www.lens.org/lens/search/scholar/list)	title:retracted OR title:retraction OR title:withdraw* FILTERS > Year Published = (*2000* - *2022*)	56,057	47,543 (84.8%)
OpenAlex	Open API (https://api.openalex.org)	Document type: https://api.openalex.org/works?filter=is_retracted:true&page=1&per-page=200 Retracted in the title: https://api.openalex.org/works?filter=title.search:retracted&page=1&per-page=200 Retraction in the title: https://api.openalex.org/works?filter=title.search:retraction&page=1&per-page=200 Withdrawn in the title: https://api.openalex.org/works?filter=title.search:withdrawn&page=1&per-page=200	69,243	57,892 (83.6%)
PubMed	Advanced Search interface (https://pubmed.ncbi.nlm.nih.gov/advanced/)	"retraction of publication"[Publication Type] OR "retracted publication"[Publication Type] OR withdraw*[Title] AND (2000:2022[pdat])	42,015	24,510 (58.3%)
Scilit	Open API (https://app.scilit.net/api/v1/solr/articles/query)	https://app.scilit.net/publications?facet={"status:["Retraction","Withdrawal"]}	53,362	52,615 (98.6%)
Scopus	Advanced document search (https://www.scopus.com/search/form.uri?display=advanced)	TITLE (retraction) OR (withdrawn) OR DOCTYPE (*tb*) AND PUBYEAR > 1999	35,219	31,634 (89.8%)
Web of Science	Advanced Search Query Builder (https://www.webofscience.com/wos/alldb/advanced-search)	(DT=(Retraction) OR DT=(Retracted Publication)) AND (PY=(2000-2022))	31,719	31,565 (99.5%)

Table 1. Data collection process carried out in each bibliographic database

There are some explanations of the query syntax and the search procedure. Retraction notices and retracted papers may both be included under the Scilit Retraction label. In WoS, withdrawals were labeled as Retraction up to 2021 (Clarivate, 2023).

After grouping all the documents that were recovered, 87,247 duplicate publications were eliminated. Data were cleansed to eliminate other types of editorial notes that some services have indexed as retractions (Table 1) (7,823; 9%), such as errata, expressions of concern, and corrigenda. Also, papers that dealt with retractions (i.e. publication ethics) or included terms such as retraction (i.e. physiology, odontology) or withdrawn (i.e. psychology, pharmacology) in the title were eliminated as false positives (6,634; 7.6%). Following this data cleaning process, 72,790 (83.4%) records were stored in a relational database.

We have distinguished retractions notices and withdrawals from retracted papers in this database. Despite the fact that some platforms connect the notices and the documents, those connections are not shown in the search results. Therefore, we have removed all the labels (i.e. retracted, withdrawn, retraction) in the title and performed a title matching between documents, assuming that the most recent publication is the notice and the oldest one is the original retracted article. However, not all the retraction notices mention the title of the retracted publication (30,266; 79.3%). In those instances, we have taken the retracted document's title directly from the abstract or the text of the retraction notice. Lastly, we have identified 34,663 (47.6%) retracted publications, 20,741 (28.5%) retraction notices and 17,387 withdrawals (23.9%).

Removing or updating the retracted article with the retraction notice is usual practice in electronic publishing, although COPE (2019) strongly advises against it. In those circumstances, retraction notices and retracted publications are the same document, with the same identifier, but with different versions that are not necessarily represented in databases. This fact happens in 40.7% (29,650) of the cases. This happens frequently with withdrawals where 83.6% (14,527) of the publications are replaced with a brief communication. However, in the case of retractions, the retraction notice replaces 43.6% (15,123) of the retracted publications. This substitution makes it more difficult to ascertain when a publication was published (particularly for articles in press) and to know the cause of a retraction.

5. Results

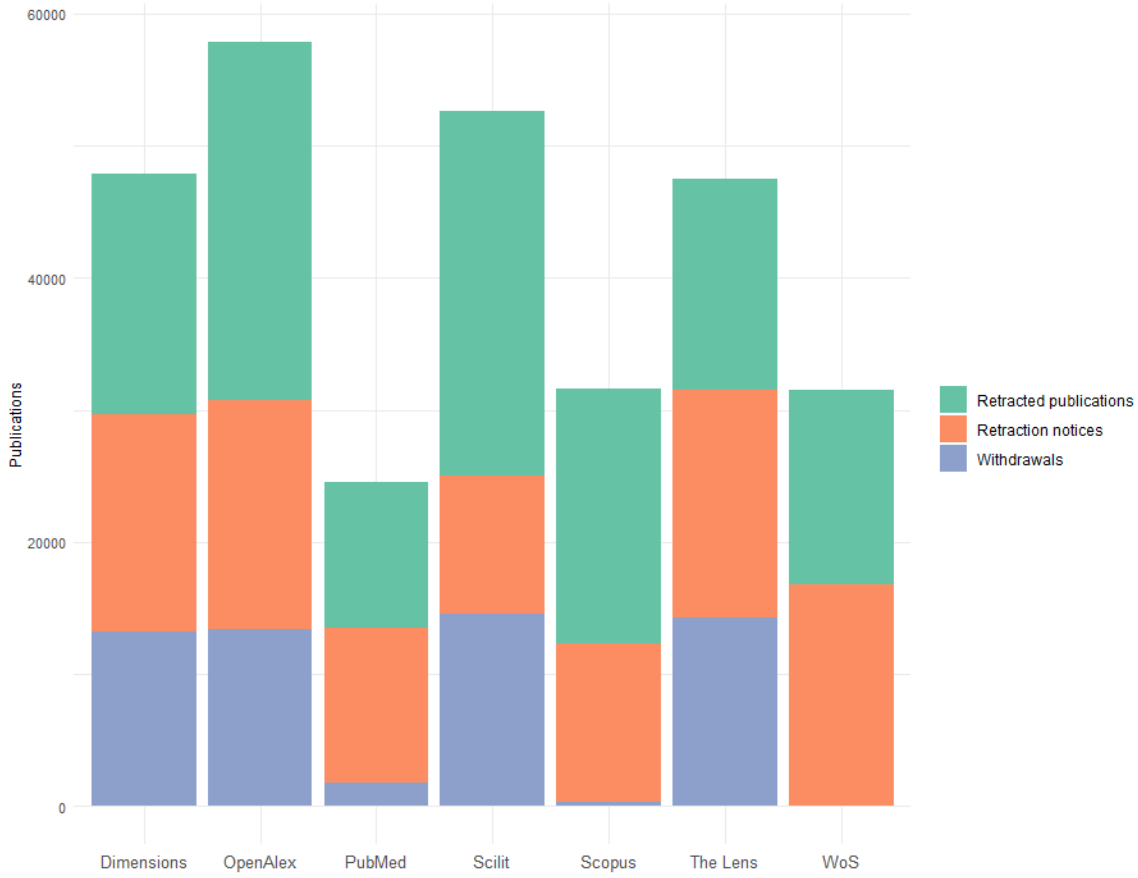

Figure 1. Coverage of retractions, withdrawals and retracted articles by database

Databases	Retraction notices	Retraction notices %	Retracted publications	Retracted publications %	Withdrawals	Withdrawals %	Total
Dimensions	16,484	34.4%	18,249	38.1%	13,180	27.5%	47,913
The Lens	17,309	36.4%	15,990	33.6%	14,244	30.0%	47,543
OpenAlex	17,329	29.9%	27,129	46.9%	13,434	23.2%	57,892
PubMed	11,724	47.8%	11,020	45.0%	1,766	7.2%	24,510
Scilit	10,470	19.9%	27,618	52.5%	14,527	27.6%	52,615
Scopus	12,025	38.0%	19,280	60.9%	329	1.0%	31,634
WoS	16,735	53.0%	14,756	46.7%	74	0.2%	31,565

Table 1. Amount and percentage of retractions, withdrawals and retracted articles in each database

Figure 1 and Table 1 show the number of retractions, withdrawals, and retracted articles that were retrieved from the chosen databases. While Scopus (31,634), WoS (31,565), and PubMed (24,510) are the bibliographic services that index fewer publications, OpenAlex (57,892) and Scilit (52,615) are the platforms that identify the greatest number of these type of publications. These coverage differences can be explained by the way in which each database is built. Scopus and WoS are selective sources that restrict their coverage to a particular set of journals, which justifies the low coverage. PubMed is a specialist database for health-related studies and medicine that exclusively indexes journals from that field of study. Scilit, Dimensions and The Lens display comparable coverage because they are recent products based on secondary sources (e.g., Crossref, PubMed, Microsoft Academic). They have a wider coverage than the earlier one because they do not restrict their indexation to particular journals or document types. OpenAlex is the bibliographic service with the widest coverage due mainly to it is based on the now-defunct Microsoft Academic, a search engine that indexed all the academic content on the Web. It is also interesting to notice that a large part of these differences is due to withdrawn publications. This document type is hardly indexed by traditional databases such as WoS (.2%), Scopus (1%), and PubMed (7.2%). Lastly, the uneven ratio in Scilit between retractions, which has the lowest percentage with 19.9%, and retracted articles, which has the second-highest percentage with 52.5%, raises the possibility that both types of publications could be misclassified.

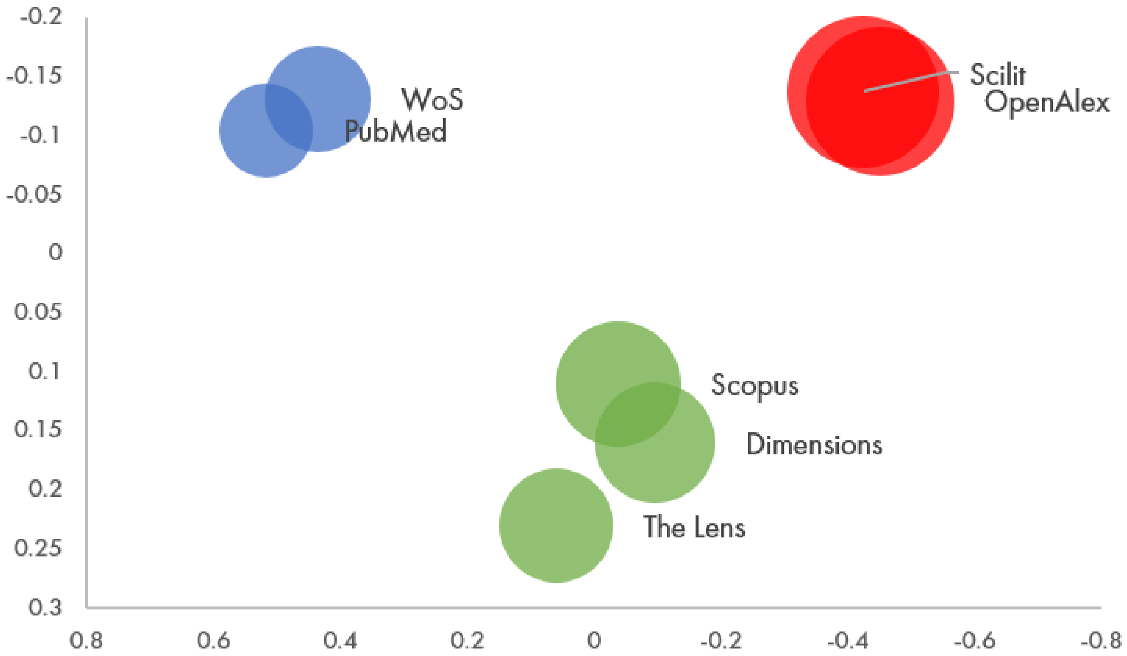

Figure 2. MDS graph showing differences between databases according to the coverage of retracted publications, retraction notices and withdrawals.

Figure 2 plots the position of the databases according to the proportion of shared publications. In other words, the quantity of papers that are jointly indexed in two databases, adjusted by the total amount of articles. A symmetrical similarity matrix was built, removing the diagonal values. The aim of this analysis is to show differences or similarities between databases. Multidimensional Scaling (MDS) was utilized to get the distance coordinates and k-means algorithm was used to find clusters. MDS is a visualization technique for displaying the information contained in a distance matrix (Kruskal & Wish, 1978). K-means is a clustering algorithm that groups elements according to the nearest mean of each cluster. Three different groups were detected: Firstly, WoS and PubMed (blue) share a great part of their records because WoS contains Medline, the principal source of PubMed. Thus, 99.5% of PubMed is indexed in WoS, and 77.3% of WoS is in PubMed. A second group (green) is shaped by Scopus, Dimensions and The Lens. The main distinction between these databases and the first group is the significant percentage of proceeding articles they index that are not included in WoS (96%) and PubMed (100%). A third group is shaped by OpenAlex and Scilit, their main characteristic is that they index the most publications without any restrictions on document type (including proceeding articles) or retraction type (including withdrawals).

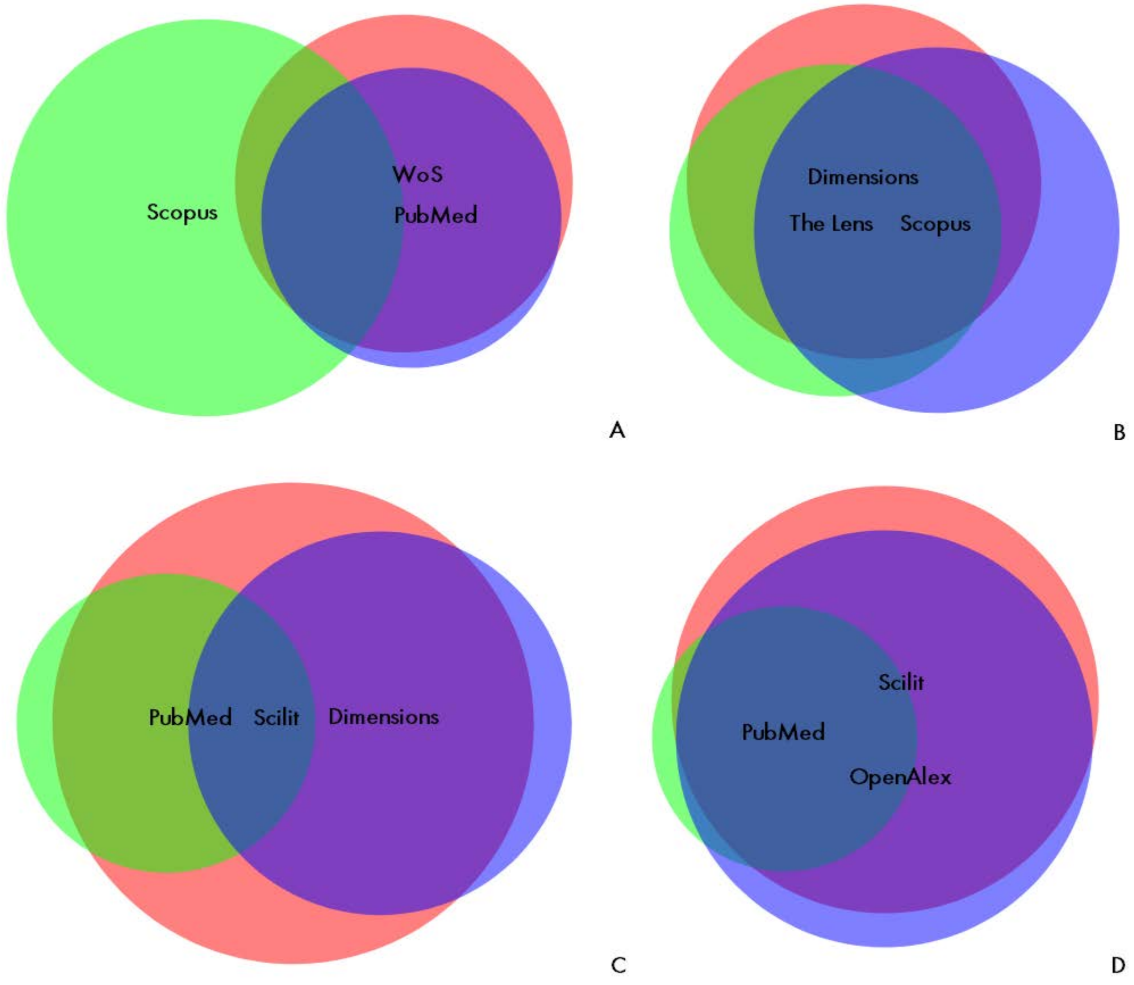

Figure 3. Some Venn diagrams showing the overlap among databases

Some Venn diagrams showing the overlap between the various databases are shown in Figure 3. The purpose of these diagrams is to clarify some specific distinctions across databases. Figure 3.a, for instance, clearly shows both the significant disparity between WoS-PubMed and Scopus, where only 22% of Scopus is indexed in PubMed and 23% in WoS, and the considerable overlap between WoS and PubMed. This is accounted for by the fact that article proceedings make up 82.6% of the missing papers in WoS and PubMed. However, a sizeable portion of WoS papers (44.2%) are not identified in Scopus, being retracted publications the most frequent (63.5%). The cause is the absence of a specific label or classification designating these missing articles as retractable. Thus, when an article is retracted, the document type is not updated to "Retracted" nor this term is added to the title. This is the reason why the document was not found using these criteria. This issue also affects Dimensions and The Lens, which would explain the closeness in Figure 2 and the overlap in Figure 3.b.

To confirm this assumption, we have searched for 2,000 retracted publications from the sample that were not retrieved in Scopus, Dimensions and The Lens. The goal is to determine the cause of these records were not retrieved. Only 34 (2.3%) of the 1,442 (72.1%) Scopus-indexed articles contained the words *retraction*, *retracted* or *withdrawn* in the title. A sizable fraction (14.9%) had the type *Erratum* rather than *Retracted*. Dimensions retrieved 1,870 (93.5%) records, but only 170 (9.1%) inserted those words in the title. In a similar vein, The Lens indexed 1,796 (89.8%), however only 124 (6.9%) publications were flagged as retracted. These results show that the proper identification and updating of these types of publications is the primary issue with these databases.

Finally, the databases with the largest coverage of retracted, retractions, and withdrawn papers are Scilit and OpenAlex (Figures 3.c and 3.d). In the case of Scilit, this database contains 88.3% of the Cochrane database's withdrawn systematic reviews. Whereas, OpenAlex (49.3%) has an important coverage of documents without DOI, along to WoS (56.8%) and The Lens (51.3%). In addition, the correct identification and labeling of this type of publications justifies their high coverage.

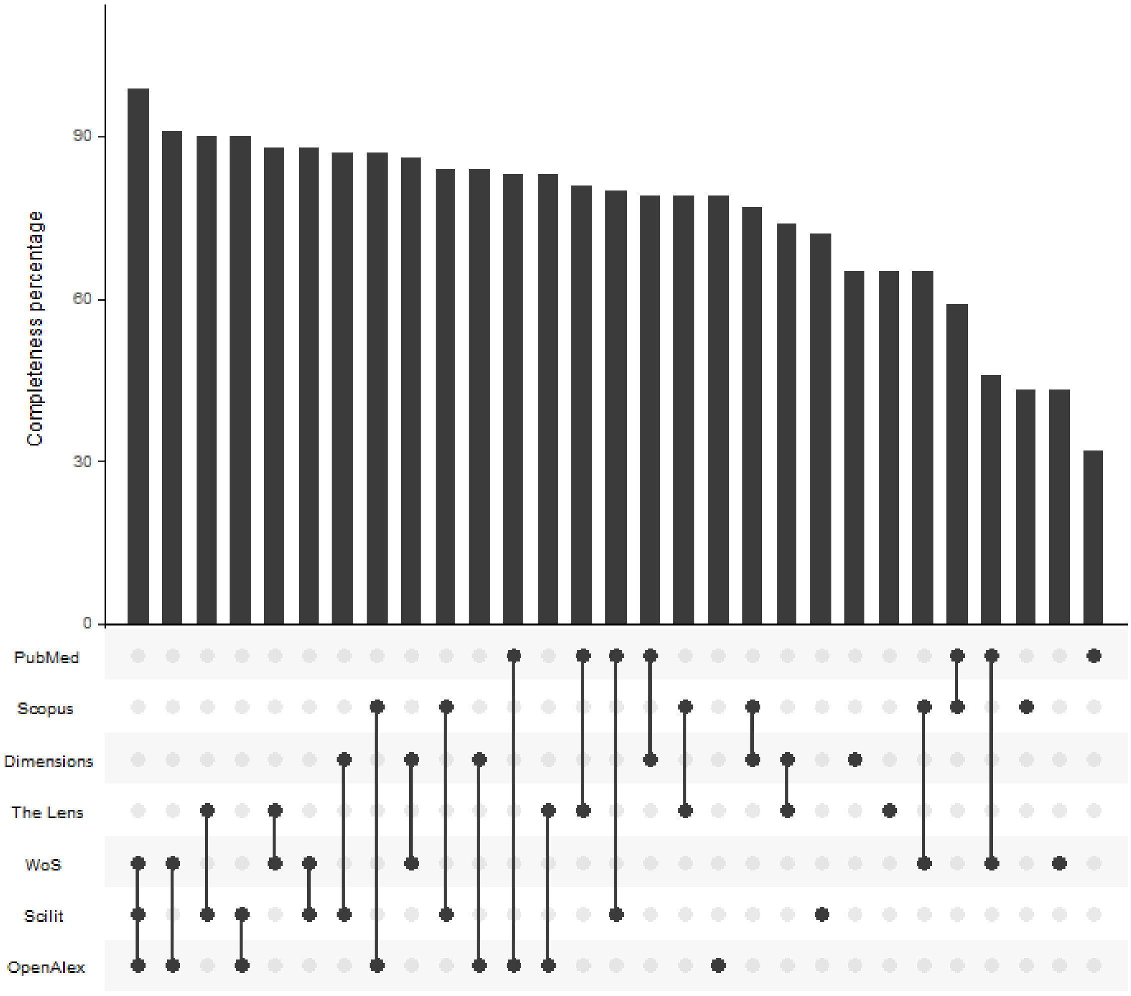

Figure 4. Combination graph showing the completeness degree according to the aggregation of different databases.

Figure 4 provides a different perspective on the overlap among databases. This combination graph shows the proportion of completeness according to each database aggregation. The purpose is to determine which database combination would yield the most comprehensive sample of retractions, withdrawals, and retracted publications. This goal is crucial to correctly tracking the development of this type of publications and choosing the best sources. The findings corroborate the high degree of source overlap, with OpenAlex, Scilit, and WoS combining to collect 99.9% of the sample. This percentage drops to 91% if OpenAlex is joint with WoS and 90.8% if the union is with Scilit. It is interesting to see that OpenAlex collects more papers on its own (79.2%) than Dimensions and Scopus (77.5%) or Dimensions and The Lens together (74.8%). Traditional sources that have been used for retraction studies also show significant low coverages. This is the case of WoS, with just 43.3% of the sample, and Scopus, with 43.4%. The combination of these last databases would actually produce (65.8%) fewer papers than The Lens (65%), Dimensions (65.6%), or Scilit (72.1%). These results demonstrate that the monitoring of a small number of sources allows us to obtain a high knowledge regarding the incidence of corrected literature, as well as open source databases such as OpenAlex are gaining ground in the scholarly environment.

6. Discussion

This study on the coverage of retracted publications, retraction notices and withdrawals in scholarly databases has showed that the first problem with retrieving this type of publications is their clear identification. Whereas common bibliographic records just need a formal description (Delgado-Quirós and Ortega, 2023), retracted publications require updating this description either including a label in the title or categorizing the record in a particular document type. This distinction is important since our study does not attempt to compare the coverage of these records, but rather how databases have identified and labeled retracted publications, and consequently they can be retrieved. This detail has evidenced that Scopus, Dimensions and The Lens do not label correctly retracted publications, making it impossible to retrieve them. This problem has been previously detected by Schneider et al. (2023). They observed that 31.7% of retracted documents in Crossref, Retraction Watch and WoS were not marked as such in Scopus. According to their investigation, Scopus incorrectly assigns 99% of retracted papers to the *tb* (Retracted) document type. This error might have occurred because a sizable fraction of retractions (15%) are marked as Erratum. In a previous study, Proescholdt and Schneider (2020) already indicated that more than 90% of Scopus' documents with the words "retracted article" in the title were not assigned to *retracted* document type.

The inclusion of withdrawals is the key component that explains coverage variations between databases. PubMed and Scopus experience significant problems on adding this type of documents, while WoS actually does not index them. The results of Schmidt (2018) also present that an important number of withdrawn publications are not labeled in PubMed (888 in 1983-2013 period), showing comparable figures with our results (1092 in 2000-2013 period). When withdrawals are taken out of the results, the discrepancies between databases are reduced, going from 22,744 for PubMed and 31,305 for Scopus to 44,458 for OpenAlex. Many of these withdrawals take place on the publisher's landing page, without retraction notice, and before articles are assigned to an issue, which may be the cause of their absences. Then, it is likely that this data is not included in the journal metadata, which serves as primary data source for PubMed, Scopus, and WoS.

To less extent, conference proceedings are another document category that significantly affects coverage. This type of publications is the factor that distinguishes PubMed and WoS from the other ones, they hardly ever index that type of publications. Overall, we can summarize that inclusive products (Dimensions, The Lens, Scilit, and OpenAlex), which do not establish formal criteria that limit the coverage of specific types of documents, have less difficulty indexing retracted literature than traditional databases (WoS, Scopus, and PubMed) based on the selection of venues. The recent results of Uppala et al. (2022), finding a low overlap between editorial notices in Crossref and PubMed, reinforce the observed differences between PubMed and Dimensions, The Lens and Scilit, databases that use Crossref as primary source (Delgado-Quirós, et al., 2024).

These findings have significant implications for studies about correction of science. Traditionally, PubMed and WoS have been used to analyze publications that have been retracted. Our findings, however, indicate that these databases are much from complete. Although those products are among the best at identifying and marking retractions, they are insufficient to follow the evolution and incidence of these problematic papers due to their scant coverage of withdrawals and proceeding articles. Our findings imply that this kind of research require the use of two or more databases, mostly non-selective sources, in order to obtain a comprehensive picture about the correction of science. In addition, this study reveals that certain databases (Scopus, Dimensions and The Lens) have issues with labeling retracted publications. This has serious implications for researchers because they could not be aware of these corrected publications, committing the mistake of inappropriate citations and validating unfound results. We recommend therefore the use of different bibliographic databases or reference managers (i.e., Zotero) to cross-check the status of the references.

We consider that in the manner that there are recommendations about how an article should be retracted (COPE, 2019), bibliographic databases should follow basic guidelines about how identify these publications and establish links between retracted documents and retraction notices and withdrawals. This is not only important to warn researchers against citing retracted publications when they conduct literature searches, but also to provide a more accurate picture about the incidence and evolution of retracted literature.

6.1. Limitations

The main limitation of this study comes from the search interfaces of the databases employed in the study. In certain instances, we have used title searches (Dimensions, The Lens) document type searches (PubMed, Scilit) or a combination of both (Scopus, OpenAlex, WoS). The way these products are searchable may have an impact on the outcomes.

7. Conclusions

Several conclusions can be drawn from this study. There are significant coverage disparities between databases. OpenAlex and Scilit are the products that identify the most withdrawn literature, while PubMed, Scopus and WoS gather the lowest percentage.

These differences may be the result of two basic causes. On the one hand, the way in which these products obtain their bibliographic metadata influences their coverage of retracted literature. For instance, the incomplete inclusion of withdrawals in PubMed, Scopus and WoS explains to great extent the coverage discrepancies between databases based on venues selection and databases based on third party sources (Dimensions, OpenAlex, Scilit and The Lens). Another crucial factor that would help to explain these disparities is the indexation of proceeding articles. On the other hand, the manner that each database labels these papers affect the coverage. The significant discrepancies between OpenAlex, Scilit, Dimensions, Scopus, and The Lens are hence the result of improper identification of retracted publications, which prevented proper retrieval.

We concluded that any study about retracted papers needs to use more than one source in order to get a trustworthy picture about these publications because of the coverage gaps between databases. The findings indicate that 99% of the sample could be retrieved using just three databases (OpenAlex, Scilit, and WoS), and 91% if only OpenAlex and WoS were combined.

8. Competing interests statement

José Luis Ortega is member of the Scilit advisory board, which made possible accessing to API documentation.

9. Funding information

This work was supported by the research project (NewSIS) "New scientific information sources: analysis and evaluation for a national scientific information system" (Ref. PID2019-106510GB-I00) funded by the Spanish State Research Agency (AEI) PN2019 and it is also part of SILICE project (Ref. PRPPDC2022-133700-I00) funded by MCIN/AEI/10.13039/501100011033 and the European Union "NextGeneration EU"/PRTR.

10. Author contribution

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by José Luis Ortega and Lorena Delgado Quirós. The first draft of the manuscript was written by José Luis Ortega and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

11.References

Aspura, M. Y. I., Noorhidawati, A., & Abrizah, A. (2018). An analysis of Malaysian retracted papers: Misconduct or mistakes?. *Scientometrics*, *115(3)*, 1315-1328. https://doi.org/10.1007/s11192-018-2720-z

Barber, M. (2021). Strengthening research integrity: The role and responsibilities of publishing. ISC Occasional paper. Retrieved June 21, 2023 https://council.science/publications/strengthening-research-integrity/

Bik, E. M., Fang, F. C., Kullas, A. L., Davis, R. J., & Casadevall, A. (2018). Analysis and correction of inappropriate image duplication: the molecular and cellular biology experience. *Molecular and cellular biology, 38*(20), e00309-18. https://doi.org/10.1128/mcb.00309-18

Bolland, M. J., Grey, A., & Avenell, A. (2022). Citation of retracted publications: A challenging problem. *Accountability in Research*, *29*(1), 18-25. https://doi.org/10.1080/08989621.2021.1886933

Budd, J. M., Sievert, M., & Schultz, T. R. (1998). Phenomena of retraction: reasons for retraction and citations to the publications. *JAMA*, *280*(3), 296-297. https://doi.org/10.1001/jama.280.3.296

Cabanac, G., Alexandre, C., Jégou, L., & Maisonobe, M. (2023, April). The Geography of Retracted Papers: Showcasing a Crossref–Dimensions–NETSCITY Pipeline for the Spatial Analysis of Bibliographic Data. In *27th International Conference on Science, Technology and Innovation Indicators (STI 2023)*. International Conference on Science, Technology and Innovation Indicators. https://dapp.orvium.io/deposits/6442fee5c93d17c257de17d2/view

Clarivate (2023). Web of Science Core Collection: Document Type Descriptions. Retrieved June 21, 2023 https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Core-Collection-Document-Type-Descriptions?language=en_US

Cokol, M., Iossifov, I., Rodriguez-Esteban, R., & Rzhetsky, A. (2007). How many scientific papers should be retracted? *EMBO reports*, *8*(5): 422-423. https://dx.doi.org/10.1038/sj.embor.7400970

Cokol, M., Ozbay, F., & Rodriguez-Esteban, R. (2008). Retraction rates are on the rise. *EMBO reports*, *9*(1), 2-2. https://doi.org/10.1038%2Fsj.embor.7401143

COPE Council (2019). COPE Retraction guidelines -- English. https://doi.org/10.24318/cope.2019.1.4

Decullier, E., Huot, L., Samson, G., & Maisonneuve, H. (2013). Visibility of retractions: a cross-sectional one-year study. *BMC research notes*, *6*(1), 1-6. https://doi.org/10.1186/1756-0500-6-238

Delgado-Quirós, L., & Ortega, J. L. (2023). Comparing publication information in seven citation indexes. In *27th International Conference on Science, Technology and Innovation Indicators (STI 2023)*. International Conference on Science, Technology and Innovation Indicators. https://dapp.orvium.io/deposits/6436c590b3340c364be5b2c7/view

Delgado-Quirós, L., Aguillo, I. F., Martín-Martín, A., Delgado López-Cózar, E., Orduña-Malea, E., & Ortega, J. L. (2024). Why are these publications missing? Uncovering the reasons behind the exclusion of documents in free-access scholarly databases. *Journal of the Association for Information Science and Technology*, *75*(1), 43-58. https://doi.org/10.1002/asi.24839

Elango, B., Kozak, M., & Rajendran, P. (2019). Analysis of retractions in Indian science. *Scientometrics*, *119*(2), 1081-1094. https://doi.org/10.1007/s11192-019-03079-y

Fanelli, D. (2013). Why growing retractions are (mostly) a good sign. *PLOS medicine*, *10*(12), e1001563. https://doi.org/10.1371/journal.pmed.1001563

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, *109*(42), 17028-17033. https://doi.org/10.1073/pnas.121224710

Grieneisen, M. L., & Zhang, M. (2012). A comprehensive survey of retracted articles from the scholarly literature. *PLOS one*, *7*(10), e44118. https://doi.org/10.1371/journal.pone.0044118

Heibi, I., & Peroni, S. (2022). A quantitative and qualitative open citation analysis of retracted articles in the humanities. *Quantitative Science Studies*, *3*(4), 953-975. https://doi.org/10.1162/qss_a_00222

Hsiao, T. K., & Schneider, J. (2021). Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine. *Quantitative Science Studies*, *2*(4), 1144-1169. https://doi.org/10.1162/qss_a_00155

Hesselmann, F., Graf, V., Schmidt, M., & Reinhart, M. (2017). The visibility of scientific misconduct: A review of the literature on retracted journal articles. *Current sociology*, *65*(6), 814-845. https://doi.org/10.1177%2F0011392116663807

Kim, S. Y., Yi, H. J., Cho, H. M., & Huh, S. (2019). How many retracted articles indexed in KoreaMed were cited 1 year after retraction notification. *Science Editing*, *6*(2), 122-127. https://doi.org/10.6087/kcse.172

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling* (No. 11). Sage. http://dx.doi.org/10.4135/9781412985130

Lu, S. F., Jin, G. Z., Uzzi, B., & Jones, B. (2013). The retraction penalty: Evidence from the Web of Science. *Scientific reports*, *3*(1), 3146. https://doi.org/10.1038/srep03146

Malkov, D., Yaqub, O., & Siepel, J. (2023). The spread of retracted research into policy literature. *Quantitative Science Studies*, *4*(1), 68-90. https://doi.org/10.1162/qss_a_00243

Nath, S. B., Marcus, S. C., & Druss, B. G. (2006). Retractions in the research literature: misconduct or mistakes?. *Medical Journal of Australia*, *185*(3), 152-154. https://doi.org/10.5694/j.1326-5377.2006.tb00504.x

Redman, B. K., Yarandi, H. N., & Merz, J. F. (2008). Empirical developments in retraction. *Journal of medical ethics*, *34*(11), 807-809. https://doi.org/10.1136/jme.2007.023069

Ortega, J. L., & Delgado-Quirós, L. (2023). How do journals deal with problematic articles? The editorial response of journals to articles commented in PubPeer. *El profesional de la Información*, 32(1). https://doi.org/10.3145/epi.2023.ene.18

Proescholdt, R., & Schneider, J. (2020). Retracted papers with inconsistent document type indexing in PubMed, Scopus, and Web of Science. Retrieved June 21, 2023 https://www.ideals.illinois.edu/items/117915

Ribeiro, M. D., & Vasconcelos, S. M. (2018). Retractions covered by Retraction Watch in the 2013–2015 period: prevalence for the most productive countries. *Scientometrics, 114(2)*, 719-734. https://doi.org/10.1007/s11192-017-2621-6

Sanderson, K. (2023). EU council's' no pay' publishing model draws mixed response. *Nature*. https://doi.org/10.1038/d41586-023-01810-7

Schneider, J., Lee, J., Zheng, H., & Salami, M. O. (2023, April). Assessing the agreement in retraction indexing across 4 multidisciplinary sources: Crossref, Retraction Watch, Scopus, and Web of Science. In *27th International Conference on Science, Technology and Innovation Indicators (STI 2023)*. International Conference on Science, Technology and Innovation Indicators. https://dapp.orvium.io/deposits/6441e5cae04dbe5586d06a5f/view

Schmidt, M. (2018). An analysis of the validity of retraction annotation in PubMed and the Web of Science. *Journal of the Association for Information Science and Technology*, *69*(2), 318-328. https://doi.org/10.1002/asi.23913

Snodgrass, G. L., & Pfeifer, M. P. (1992). The characteristics of medical retraction notices. *Bulletin of the Medical Library Association*, *80*(4), 328.

Stricker, J., & Günther, A. (2019). Scientific misconduct in psychology. *Zeitschrift für Psychologie*. https://doi.org/10.1027/2151-2604/a000356

Suelzer, E. M., Deal, J., Hanus, K. L., Ruggeri, B., Sieracki, R., & Witkowski, E. (2019). Assessment of citations of the retracted article by Wakefield et al with fraudulent claims of an association between vaccination and autism. *JAMA network open*, *2*(11), e1915552-e1915552. https://doi.org/10.1001/jamanetworkopen.2019.15552

Trikalinos, N. A., Evangelou, E., & Ioannidis, J. P. (2008). Falsified papers in high-impact journals were slow to retract and indistinguishable from nonfraudulent papers. *Journal of clinical epidemiology*, *61*(5), 464-470. https://doi.org/10.1016/j.jclinepi.2007.11.019

Uppala, A., Rosati, D., Nicholson, J. M., Mordaunt, M., Grabitz, P., & Rife, S. C. (2022). Title detection: a novel approach to automatically finding retractions and other editorial notices in the scholarly literature. *arXiv preprint arXiv:2210.09553*.

Vuong, Q. H. (2020). Reform retractions to make them more transparent. *Nature*, 582, 149 https://doi.org/10.1038/d41586-020-01694-x