An Analysis of Availability and Implications of Unlabeled Retracted Articles on Sci-Hub

Biju V. V.[a,b,*], Sanjo Jose [b,c], Franklin J.[d,], and Jasimudeen S.[e,f.]

[a*] *Librarian, Sacred Heart College (Autonomous), Thevara, Kochi, Kerala, India 682 013. https://orcid.org/0000-0001-5075-0385; bijuvv@shcollege.ac.in ; [b] Research Scholar, Bishop Heber College (Autonomous), Tiruchirappalli, (Affiliated to Bharathidasan University), Tamil Nadu, India 620 017; [c] Librarian, St. Thomas College (Autonomous), Thrissur, Kerala, India - 680 001, https://orcid.org/0000-0002-9530-0393; [d] Assistant Professor, PG and Research Department of Library and Information Science, Bishop Heber College (Autonomous), Tiruchirappalli, (Affiliated to Bharathidasan University), Tamil Nadu, India, 620 017. https://orcid.org/0000-0001-6792-9752; [e] Librarian, St. Stephen's College, Uzhavoor, Kottayam, Kerala, India – 686 634, https://orcid.org/0000-0002-1451-7882; [f] Research Ethics Committee Member, National Institute of Physical Medicine and Rehabilitation (NIPMR), Thrissur, Kerala, India -680 683.*

An Analysis of Availability and Implications of Unlabeled Retracted Articles on Sci-Hub

Researchers are increasingly accessing scientific articles through unauthorized websites like Sci-Hub. Sci-Hub contains retracted articles, including those which are not labelled as retracted, and this is a potential threat to academic research. This study analyses the extent of the availability of retracted articles within the Sci-Hub, particularly focusing on the presence of unlabeled retracted articles (URA) which may inadvertently be used in subsequent research, thus propagating flawed findings. The authors identified 16,925 English-language research articles retracted between 2003 and 2022 indexed in the Web of Science and Scopus databases. These articles were cross-checked with Sci-Hub to ascertain whether they were appropriately labelled as retracted. The investigation revealed that 84.83% of the retracted articles available on Sci-Hub do not have any indication of their retracted status. These URA could potentially be reused by researchers, unaware of their retracted status. The availability of URA in the field of health sciences is particularly high, which indicates a significant risk of their unintended use and further citation in future research. This study underscores the crucial need for stringent implementation of regulatory measures suggested by Committee on Publication Ethics (COPE) or newly published National Information Standards Organization (NISO) recommendations. Users of Sci-Hub should cross check the validity of articles downloaded from it with credible sources to prevent the inadvertent citation of retracted articles.

Keywords: Academic Publishing; Research Integrity; Post-Retraction Citations; Article Retraction; Black Open Access; Scholarly Communication.

Subject classification codes: Library Science

Introduction

The conventional system of academic publishing necessitates subscription or an institutional license to access research outputs. Although Open Access routes like Gold and Green came into existence as alternatives for such publishing platforms, a substantial portion of research articles remains behind paywalls. Researchers or scientists require immediate access to pertinent information as delayed information can impede the researcher's progress (Braha, 2017). Shadow libraries like Sci-Hub emerged to address this issue, which was subsequently termed as Black Open Access (Björk, 2017). Sci-Hub stands as the world's busiest shadow library for accessing scientific articles and the easiest way to find out the PDF version of a desired article behind paywall with a single click (Faust, 2016). Sci-Hub espouses the belief that

scientific knowledge should be universally accessible to all regardless of an individual's fiscal resources, social stature, or geographic locale. In this capacity, it aligns with the Open Access movement, advocating for freely accessible research results - a stark contrast to conventional subscription models (Himmelstein et al., 2018).

Sci-Hub

Most of the research articles are published behind paywalls rendering them inaccessible to many of the researchers worldwide, particularly those in low-income countries (Kirsop & Chan, 2005). Various published studies have revealed that researchers worldwide are utilizing Sci-Hub for downloading research articles, without any discrimination of being developed or developing countries (Amin et al., 2021; Behboudi et al., 2021; Bohannon, 2016; Elbakyan, 2024b; Geng et al., 2022; Greshake, 2017; Owens, 2022; Sagemüller et al., 2021; Wan, 2022). Table 1 is a replication of the subject wise availability of 88 million articles on Sci-Hub as furnished on its website on 10th April 2024. [Table 1 near here]

Retraction

Committee on Publication Ethics (Barbour et al., 2009; Kleinert, 2009; Wager et al., 2010) states retraction as a self-correction mechanism within the scientific domain. The act of retraction aims to invalidate the published material and alert readers to refrain from any further use or citation of a retracted article (Dal-Ré & Ayuso, 2020; Kohl & Faggion, 2024; Xu & Hu, 2022). Retraction notices serve to specify the reasons for retraction thereby cautioning the readers of the issues associated with it (Tripathi et al., 2019).

Concepts

Version of Records (VoR): According to Crossref, (n.d.) version of record is the typeset, copyedited, and published version of an article. These are the versions which one can access on publishers' webpage usually as full text.

Labelled Retracted Articles (LRA): These are updated versions of records or any other version of article that include a heading or watermark indicating its retraction status.

Unlabeled Retracted Version of Records (URVoR): This refers to versions of records of retracted articles that lack any form of labelling, such as watermarks or headings, to indicate their retraction status. This absence of labelling can create a misleading impression that the article remains valid despite being retracted.

Unlabeled Retracted Articles (URA): This refers to those articles which have been retracted but do not bear any kind of labelling like watermarks or headings to indicate that they have been retracted. Apart from unlabeled retracted version of records (URVoR), URA encompasses various versions of articles including preprints, accepted manuscripts, uncorrected proofs, articles in press, and corrected proofs.

Retracted articles available on Sci-Hub

Information seekers who send a query to the Sci-Hub database will receive the full text of the article if available, usually in the form of a PDF file. However, some of the journal articles accessed by the researchers might have been retracted. While some of these downloaded articles may be labeled as retracted within their PDF file, many lack any indication of their retraction status when accessed from Sci-Hub. Boudry et al., (2023) identified the presence of non-retracted versions of the retracted article within the Sci-Hub.

The process of announcement of the retraction status of an article involves several sequential phases. The duration of retraction announcements can vary across different subjects and publishers (Elango, 2021; Feng et al., 2020; King et al., 2018). It is possible for the original article to be archived in Sci-Hub before the retraction of a scholarly publication. In the context of Sci-Hub, it is the utilization of DOIs that the indexing of articles mostly relies on. Once an article has been indexed using a specific DOI, replacement with its subsequent versions seems to be not happening due to the identical DOIs across the versions. Unaware of its retracted status, users may access and potentially cite such URA in subsequent articles, assuming them to be valid. It was also observed that retraction notices with a different DOI other than the original article were indexed in full text in Sci-Hub.

There may be three possibilities for accessing articles from Sci-Hub as URA by the information seekers unknowingly.

1. Some users directly access articles from Sci-Hub even without visiting the article's webpage, when they are in possession of Digital Object Identifiers (DOIs) obtained from sources like reference lists.

2. Another category is those who try to obtain papers from the publisher websites or databases but rely on Sci-Hub due to the paywalls. In these cases, the publishers might not properly mention the article's retraction status in the webpage as proved by Bakker et al., (2024); Suelzer et al., (2021).

3. If the end user is aware that the journal they require is not subscribed to or not accessible to them, they may resort to Sci-Hub, even without visiting the journal webpage. In such a case, even if the webpage of the paper has a clear labelling that it is retracted, the user will not have a chance to see the retraction statement.

Problem statement

Researchers who use articles from Sci-Hub without verifying their retraction status may unknowingly incorporate retracted or withdrawn articles in their future work, thereby undermining the validity and reliability of their research. In this article, the authors aim to determine the extent of URA available through Sci-Hub.

Scope of the study

The presence of retracted papers within a journal serves as an indicator of the journal's commitment to maintaining scientific integrity (Fanelli, 2013). However, the presence of URA in any platform may compromise the quality of scientific discourse. The current study is done from the perspective of a Sci-Hub user, focusing on the accessibility and extent of URA on Sci-Hub. The authors also suggest potential measures to address future issues that may arise from their presence.

Researchers around the globe prefer WOS and Scopus for finding their review/related articles. Using the DOIs obtained from these databases, there are chances that the researchers may proceed to obtain full text from Sci-Hub. This may lead to the inadvertent use of URA in their works which pollute science. So, the authors chose to study the extent of availability of articles indexed in WoS and Scopus as URA in Sci-Hub.

Methodology

Web of Science (WoS) and Scopus databases were utilized to identify retracted articles indexed within them. Bibliographic details of these articles were downloaded from both databases and articles thus obtained were considered for this study. The duration of this study spans twenty years from 2003 to 2022. Scopus was searched on 27th July 2023 using the query *ALL ("Retracted Article" OR "Retraction Note" OR "Retraction Notice" OR "Withdrawn Article" OR "Retraction of Volume" OR "Notice of Withdrawal")*, and 28094 articles were retrieved. Similarly, the WoS was searched on 29th July 2023 using the query *TS=("Retracted Article" OR "Retraction Notice" OR "withdrawn article" OR "notice of withdrawal" OR "Retraction of vol" OR "retraction of" OR "retracted of") OR TI=("Retracted Article" OR "Retraction Notice" OR "withdraw article" OR "notice of withdrawal" OR "Retraction of vol" OR "retraction of" OR "retracted of")*, For this query, the WoS listed bibliographic details of 25472 retracted articles.

The records from each database were downloaded year-wise and compiled as 40 datasets and further merged year-wise reducing the number of datasets to 20. These datasets were screened for various elimination criteria as follows. Articles in English language only were considered for the study. Research articles about retraction in academic publishing and with alternate meanings like retraction in tissue culture, canine retraction, endothelial retraction, membrane retraction, retraction in business cycle, etc. were removed. Records on retractions in book chapters and conference proceedings were removed along with Cochrane reviews and chemical structures in Scopus. Multiple entries for the same article were removed to retain unique records. Some records containing errors and no DOIs were also removed. All twenty files were merged to form a single dataset containing 22060 records. As the publication and retraction happen usually in different years, the merged dataset contained both retracted articles and their corresponding retraction notices. These were further identified manually using MS Excel by comparing their DOIs and titles. 5135 such duplicate entries were located, and they were removed thereby the final dataset containing records of 16925 unique articles published by 354 different publishers. Each article was searched on Sci-Hub

using the DOI of the original article to determine their availability. After checking articles accessed through Sci-Hub, the downloaded PDF files were classified based on the watermark on them. The classification was as follows: (1) Articles with a retraction notice attached as a separate page or paragraph; (2) Articles with a watermark "retracted"; (3) Articles with a retraction mentioned at the top of every page; (4) Articles with no retraction indication as above; (5) Articles available not as VoR, but in different versions like preprints, accepted manuscripts, corrected proofs, uncorrected proofs and articles in press, which obviously will not have retraction label or marking; (6) Articles not available on Sci-Hub for download. An article was categorized as LRA if it met any of the criteria from 1 to 3. Articles that did not mention retraction and lacked any watermark indicating retraction status as in 4th and 5th categories were classified as URA (Table 2). Files in the 6th case were noted as 'article not available in Sci-Hub.' The process of gathering data was done initially by the first author and subsequently cross-checked by the second author to ensure accuracy and consistency.
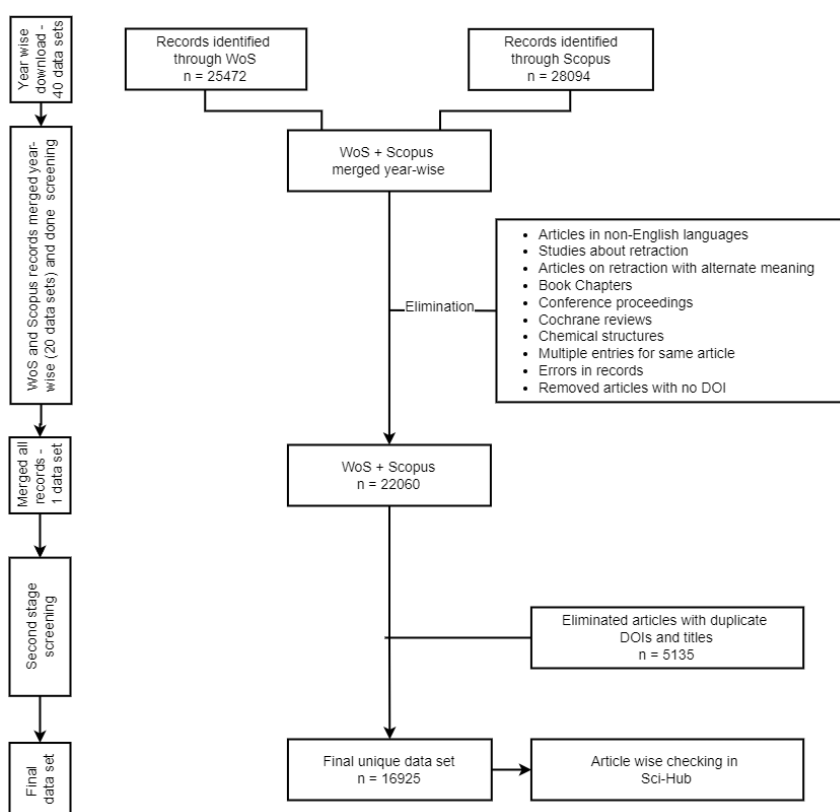
The records for each article downloaded from WoS and Scopus contained keywords. These keywords were further mapped into broader, subject areas such as Agriculture, Arts and Humanities, Engineering, Environmental Sciences (EVS), Geology, Health Sciences, ICT, Life Sciences, Management, Mathematics, Physical Sciences, and Social Sciences.

Results

The articles are categorized into eight distinct versions: URVoR (65.09%), Article not available (15.75%), LRA (12,78%), Preprints (3.27%), Accepted Manuscripts (2.61%),

Uncorrected Proofs (0.38%), Articles in Press (0.11%) and Corrected Proofs (0.01%). This distribution, as described in Table 2, reflects the dynamic nature of the versions of the indexed documents in Sci-Hub, encompassing stages from initial submission to post-publication corrections. [Table 2 and Figure 1 near here]

The analysis revealed that a significant 84.25% (16925 – 2666 =14,259) of articles are available in Sci-Hub. However, only 15.17% (n=2,163) of the available 14,259 articles indicate their retraction status (LRA), while a substantial 84.83% (12,096) articles do not mention their retraction status (URA).

Table 3 presents the distribution of retracted articles based on the year of their retraction. This table shows the growth of the articles indexed in each category. The categorization, detailed in Table 3, is based on the watermarks of the available articles in Sci-Hub, as mentioned in the methodology. Figure 2 depicts the growth graphically. [Table 3 and Figure 2 near here]

All the articles were grouped under their broad subject categories for this study, as shown in Table 4 and column 4 provides the actual number of URA available in Sci-Hub under each subject category, after excluding LRA and articles not available in Sci-Hub. Figure 3 represents the data graphically. [Table 4 and Figure 3 near here]

Health sciences, engineering, and life sciences exhibit the highest numbers of retracted articles, with many accessible in Sci-Hub without any evidence of retractions, raising concerns about potential misinformation (Table 5). Despite having fewer retractions, disciplines such as geology, and mathematics also display a notable proportion of articles available as URA. Most of the retracted health sciences articles are available in Sci-Hub (81.62%) among which 68.76% are URA. Chronological distribution of the retracted articles in health sciences is shown in Table 5. [Table 4 and 5 near here]

Based on information available from the downloaded records, the authors identified the publishers of journals that retracted articles and grouped the imprints under the respective main publishers. A total of 354 publishers were identified and were ranked based on the number of retracted articles; the top 50 are listed in Table 6 along with the number of articles not available, LRA and URA. [Table 6 in appendix] Table 7 lists the top 10 publishers with their article retraction status. In this table "RA" represents the number of retracted articles identified. [Table 7 near here]

Discussion

Available versions of articles in Sci-Hub

Since various versions of retracted articles are available as URA, common information seekers may not suspect any threat when accessing these articles from Sci-Hub. This can lead to further citations and the incorporation of the already declared erroneous results, methodology, or data in future research. Numerous studies have identified that retracted articles continue to receive citations even after formal retraction (Ajiferuke & Adekannbi, 2020; Bolboacă et al., 2019). Most post-retraction citations are examined and found to be positive in nature (Bar-Ilan & Halevi, 2017). Sci-Hub might have contributed to this phenomenon by giving access to URA. The impact of post-retraction citations extends beyond academia as they can perpetuate false findings and lend undue legitimacy to flawed research, thereby influencing a wider audience (Candal-Pedreira et al., 2020).

Table 3 highlights the availability of retracted articles in Sci-Hub chronologically across various published versions without indicating their retraction status. URVoR consists of the majority of the versions available with 65.09% (n=11017). URVoR has the highest possibility of usage among the URA as they appear to be legitimate published versions without any kind of labeling. Preprints of retracted articles have seen an increase since 2017, reflecting their widespread early dissemination. Accepted manuscripts of retracted articles have also grown considerably since 2016, indicating that even in early stages of publication, these articles are accessible on Sci-Hub. Uncorrected proofs, corrected proofs and articles in press of retracted articles are also available in Sci-Hub even though their number is relatively few. This visibility trend of URA underscores a critical issue in academic publishing, where retracted articles continue to be accessed and disseminated through platforms like Sci-Hub, potentially misleading end users who may cite these articles under the impression that they are not retracted.

Before and after Sci-Hub

While it is true that the establishment of Sci-Hub in 2011 (Himmelstein et al., 2018) has likely contributed to the increased availability of URA, it is important to note that articles retracted before 2011 are also available on the platform as URA (Table 3). Before 2011, the number of URA in Sci-Hub ranged from 24 in 2003 to a peak of 151 in 2011. This indicates that the dissemination of URA on Sci-Hub is not solely due to the platform itself; rather, it also results from these URA being available in the source databases at the time they were indexed. Post-2011, the numbers of URA escalated dramatically, with 282 articles in 2012 rising to an alarming 2,178 by 2022. URA grew from 24 in 2003 to over two thousand in just under two decades, almost 90 times.

Subject categories

Table 4 illustrates that the present study aligns with previous research confirming that retractions are prevalent across various disciplines (Wager & Williams, 2011). Almost half (49.95%) of the retracted articles identified by the authors are from health sciences. The articles from the disciplines of life sciences (16.13%), engineering (9.12%) and physical sciences (8.7%) contribute a larger share of retracted articles.

Engineering (81.92%) has the highest proportion of URA against the retracted articles identified. This reflects a significant exposure of potentially flawed research in a field critical to technological and infrastructural development. Similarly, physical sciences and life sciences demonstrate high percentages of 74.39% and 73.55% respectively, indicating a strong presence of retracted articles in disciplines that profoundly impact scientific progress. Table 4 underscores a troubling trend, where a substantial proportion of retracted articles, particularly in high-stakes fields like health sciences (68.76%) and life sciences (73.55%), remain easily accessible as URA.

Articles in health sciences

The health sciences sector exhibited a steady growth in retracted articles each year (Table 5). The authors identified 8454 unique retracted articles in health sciences from both indexing databases, among which 1554 were not available in Sci-Hub and 1087 articles downloaded from Sci-Hub were labelled. A substantial portion of these articles (n=5813), despite their retraction, remain available without any indication of their retracted status, peaking at 1181 instances in 2021 alone.

Sci-Hub declares that its goal is to provide free and unrestricted access to all scientific knowledge ever published in journal or book form. It also claims that it has addressed the issue of limited access to medical information in developing countries, a challenge acknowledged but till date unresolved by the World Health Organization. In doing so, it claims to assist "millions of ill people and health professionals in the dark" (Elbakyan, 2024a). Medicine appears to have the largest number of Sci-Hub supporters, as evident in many studies done by Culquichicón et al., (2022)., Faust, (2016), and Mejia et al., (2017). Retracted articles in health sciences can contain flawed methodologies or erroneous conclusions, which could compromise the validity of the research and thus cause harm to the patients and their lives (Steen, 2011). So, the researchers using articles downloaded from Sci-Hub should ensure that the articles they cite have not been retracted.

Publishers and retraction

Table 6 ranks publishers by the number of articles retracted (RA) by them and provides details of the availability of these retracted articles on Sci-Hub, including the mention of retractions, and the presence of URA. The number of retractions varies widely among publishers and doesn't show a clear correlation with the number of retracted articles indexed in Sci-Hub. Table 6 lists 50 publishers with the highest number of retractions.

Springer-Nature and Elsevier top the list showcasing their significant footprint in academic publishing. The high volume of retractions from these publishers reflects both their extensive publication output and potentially stringent retraction policies aimed at maintaining research integrity.

Springer-Nature tops the list with 3534 retracted articles of which 85.43% (n=3019) are available on Sci-Hub as URA. Elsevier has 2,527 URA out of 3,263 retracted articles, representing 77.44%. Wiley exhibits 1,507 URA out of 2,175, which is 69.29% of their total retractions. Taylor and Francis have 728 URA out of 968 retracted articles, translating to 75.21%. Wiley has a higher percentage of retracted articles not available on Sci-Hub (18.85%) compared to the other top four publishers.

All retracted articles are not indexed in Sci-Hub. The non-availability of retracted articles in Sci-Hub may be attributed to its indexing capabilities or to the security policies of individual journals' or publishers' that prevent articles from being indexed on the platform. For instance, Verduci Publisher has a high percentage (98.50%) of retracted articles not available on Sci-Hub, despite having 267 retractions, which may be due to their open access nature.

Top 10 publishers – their retracted articles identified and availability in Sci-Hub The comprehensive analysis of retracted articles across the top 10 publishers (Table 7) and their availability in Sci-Hub provides critical insights into the dynamics of scientific publishing and retraction practices. This table reveals the publisher wise indexing pattern of Sci-Hub chronologically. The presence of retracted articles indexed on Sci-Hub before 2011, despite the platform only being established that year, indicates that Sci-Hub has retrospectively added these articles to its database. For instance, Elsevier retracted 13 articles in 2003; among them 8 are indexed in Sci-Hub as URA. These 8 articles should have been labelled as retracted in 2003 itself, but the unlabeled versions of these articles are available in Sci-Hub which was established only in 2011 is a paradox. Such retrospective indexing often results in URA, as Sci-Hub does not always indicate their retraction status. Similar trends are observed with other publishers like Springer-Nature, Wiley, Taylor & Francis, etc. where articles published before 2011 are indexed on Sci-Hub without retraction notices or indications of retraction.

Publishers' responsibility

The Committee on Publication Ethics (COPE, 2019) and National Information Standards Organization, (2024) have published their guidelines. These guidelines emphasize the importance of marking of retracted publications and making their associated retraction notices readily accessible to effectively convey the reasons behind the retraction to prevent propagation of erroneous data and to safeguard the reliability of subsequent   research (Kleinert, 2009). Many publishers do not fully adhere to these practices. Studies have indicated that there are cases where retraction notices lack to convey adequate information or fail to meet the standards specified by COPE (Bozzo et al., 2017; Taros et al., 2023). The authors suggest that apart from the notices, the original retracted article also should be available in open access with adequate markings like visible watermark and page headings. If publishers strictly adhere to the COPE and

Communication of Retractions, Removals, and Expressions of Concern (CREC) guidelines issued by NISO, the end user may not download these articles from an alternative database like Sci-Hub, where retracted articles are available as URA. Abiding by such guidelines will at least minimize the chances to download the URA.

Users' responsibility

It is crucial for the scientific community to disseminate information and increase awareness among the academic community regarding the concerns associated with retracted works (Rosenkrantz, 2016; Tripathi et al., 2018). Information seekers using articles from unauthorized sources should exercise caution by verifying the legitimacy and credibility of the articles from the journal's official webpage or other credible source like Retraction Watch Database before utilizing the concepts or data presented within. It is important to be aware that all versions of articles available on Sci-Hub are not credible or represent the latest versions. Readers should keep this concept in mind that once a document is indexed in Sci-Hub, subsequent updates- such as new versions or corrections or retractions- do not replace the originally indexed version, which remains unchanged. Reference management software packages like Zotero and EndNote provide an alerting facility while citing retracted articles. These tools will help authors of scientific articles to recognize retracted articles to a certain extent.

Limitations

This study acknowledges some limitations, such as the exclusion of non-English articles, period of study and the dependence on WoS and Scopus only for identifying retracted articles. Retraction Watch Database indexed more articles than WoS and Scopus combined together. But the database was not used in this study, as it is not used by the researchers for finding their review/relates studies. Further, this study does not attempt to address any ethical considerations or copyright violations related to Sci-Hub.

Conclusion

Even though Sci-Hub plays a pivotal role in democratizing access to scientific knowledge, it also brings challenges that require greater vigilance and responsibility from both researchers and publishers. This study assessed Sci-Hub and found the presence of URA and pointed out the potential risks which may happen to the research ecosystems due to these URA. Out of 16925 unique articles identified from WoS and Scopus, a majority are available in Sci-Hub. Among those accessible articles 84.83% (12,096 articles) were found to be URA. Although all disciplines have a substantial presence of URA in Sci-Hub, health sciences researchers should take an extra-cautionary approach when accessing articles from Sci-Hub as 68.76% of retracted articles in these disciplines are available as URA. While conceiving an experiment, there are chances for a researcher to adopt the methodologies from a URA which might have been accessed from Sci-Hub. If such an article is relied by a researcher, it will lead to the loss of time, money and man-hours and in extreme cases lives too. This situation poses a significant risk, particularly in health-related research, where the implications of

utilizing flawed studies can have direct and profound impacts on public health policies and personal health decisions.

Publishers and databases should follow COPE and the newly formulated NISO guidelines to handle retracted literature by providing freely available retraction notices and full text of clearly watermarked retracted articles supplemented with maximum available information. Such retracted articles searchable through any search engines and easily downloadable may largely reduce the further usage and positive citations of URA from platforms like Sci-Hub.

Chances of updating the black open-access sites with the latest versions of the articles seem to be little. The presence of URA in Sci-Hub highlights an urgent need for developing an improved mechanism to prevent these articles from influencing further research and reduce the chances of post-retraction citations. To safeguard the integrity of scientific research, it is essential that the academic community adopts robust methods for tagging and tracking retracted articles. Increased awareness and proactive measures are essential to prevent the dissemination of invalidated research and to maintain the credibility of scientific communication. This study calls for a concerted effort from all stakeholders to enhance transparency and ensure that accurate, reliable information serves as the foundation for future scientific endeavors.

Abbreviations used

- DOI – Digital Object Identifier
- EVS – Environmental Science
- ICT – Information & Communication Technology
- LRA – Labeled Retracted Article
- OUP – Oxford University Press
- RA – Retracted Article
- T&F – Taylor and Francis
- URA – Unlabeled Retracted Article
- URVoR – Unlabeled Retracted Version of Record
- VoR – Version of Record
- WK – Wolters Kluwer
- WoS – Web of Science

Appendix
Table 6: https://doi.org/10.6084/m9.figshare.27445167.v1

Acknowledgements

Disclosure statement

Ethical statement

This article does not contain any studies with human participants performed by any of the authors.

Statement of author contributions

BVV conceived the idea, developed the research design, collected, coded, and analyzed the data, and drafted the initial version of the manuscript. SJ cross-checked the data. SJ and JS contributed to the development of the research design and collaborated with BVV in preparing the later versions of the manuscript. FJ supervised all research activities and provided guidance throughout the project.

ORCID

Biju V. V. https://orcid.org/0000-0001-5075-0385
Sanjo Jose https://orcid.org/0000-0002-9530-0393
Franklin J. https://orcid.org/0000-0001-6792-9752
Jasimudeen S. https://orcid.org/0000-0002-1451-7882

References

Ajiferuke, I., & Adekannbi, J. O. (2020). Correction and retraction practices in library and information science journals. *Journal of Librarianship and Information Science*, *52*(1), 169–183. https://doi.org/10.1177/0961000618785408

Amin, R., Ayoub, A., Amin, S., & Wani, Z. A. (2021). Toll-based access vs pirate access: a webometric study of academic publishers. *Digital Library Perspectives*, *38*(1). https://doi.org/10.1108/DLP-12-2020-0127

Bakker, C. J., Reardon, E. E., Brown, S. J., Theis-Mahon, N., Schroter, S., Bouter, L., & Zeegers, M. P. (2024). Identification of retracted publications and completeness of retraction notices in public health. *Journal of Clinical Epidemiology*, *173*, 111427. https://doi.org/10.1016/j.jclinepi.2024.111427

Barbour, V., Kleinert, S., Wager, E., & Yentis, S. (2009). *Guidelines for retracting articles*. https://doi.org/10.24318/cope.2019.1.4

Bar-Ilan, J., & Halevi, G. (2017). Post retraction citations in context: a case study. *Scientometrics*, *113*(1), 547–565. https://doi.org/10.1007/s11192-017-2242-0

Behboudi, E., Shamsi, A., & Bueno de la Fuente, G. (2021). The black crow of science and its impact: analyzing Sci-Hub use with Google Trends. *Library Hi Tech*, *39*(4). https://doi.org/10.1108/LHT-04-2020-0105

Björk, B.-C. (2017). Gold, green, and black open access. *Learned Publishing*, *30*(2), 173–175. https://doi.org/10.1002/leap.1096

Bohannon, J. (2016). Who's downloading pirated papers? Everyone. *Science*, *352*(6285). https://doi.org/10.1126/science.352.6285.508

Bolboacă, S. D., Buhai, D.-V., Aluaș, M., & Bulboacă, A. E. (2019). Post retraction citations among manuscripts reporting a radiology-imaging diagnostic method. *PLOS ONE*, *14*(6), e0217918. https://doi.org/10.1371/journal.pone.0217918

Boudry, C., Howard, K., & Mouriaux, F. (2023). Poor visibility of retracted articles: a problem that should no longer be ignored. *BMJ*. https://doi.org/10.1136/bmj-2022-072929

Bozzo, A., Bali, K., Evaniew, N., & Ghert, M. (2017). Retractions in cancer research: a systematic survey. *Research Integrity and Peer Review*. https://doi.org/10.1186/s41073-017-0031-1

Braha, J. (2017). Science communication at scientific societies. In *Seminars in Cell and Developmental Biology* (Vol. 70). https://doi.org/10.1016/j.semcdb.2017.05.010

Candal-Pedreira, C., Ruano-Ravina, A., Fernández, E., Ramos, J., Campos-Varela, I., & Pérez-Ríos, M. (2020). Does retraction after misconduct have an impact on citations? A pre-post study. *BMJ Global Health*, *5*(11), 1–7. https://doi.org/10.1136/bmjgh-2020-003719

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

Crossref. (2024). *Version control, corrections, and retractions - Crossref*. https://www.crossref.org/documentation/principles-practices/best-practices/versioning/

Culquichicón, C., Valladares-Garrido, M. J., Mejia, C. R., Rojas-Alvarado, A. B., Araujo-Chumacero, M. M., Córdova-Agurto, J. S., Fiestas, J., & Rojas-Vilar, F. J. (2022). Factors associated with producing a scientific publication during medical training: Evidence from a cross-sectional study of 40 medical schools in Latin America. *F1000Research*, *9*. https://doi.org/10.12688/f1000research.26596.2

Dal-Ré, R., & Ayuso, C. (2020). For how long and with what relevance do genetics articles retracted due to research misconduct remain active in the scientific literature. *Accountability in Research*, *00*(00), 1–17. https://doi.org/10.1080/08989621.2020.1835479

Elango, B. (2021). Retracted articles in the biomedical literature from Indian authors. *Scientometrics*, *126*(5), 3965–3981. https://doi.org/10.1007/s11192-021-03895-1

Elbakyan, A. (2024a). *Sci-Hub*. https://sci-hub.se/medicine

Elbakyan, A. (2024b). *Sci-Hub usage statistics*. https://sci-hub.se/stats

Fanelli, D. (2013). Why Growing Retractions Are (Mostly) a Good Sign. *PLoS Medicine*, *10*(12), e1001563. https://doi.org/10.1371/journal.pmed.1001563

Faust, J. S. (2016). Sci-Hub: A Solution to the Problem of Paywalls, or Merely a Diagnosis of a Broken System? In *Annals of Emergency Medicine* (Vol. 68, Issue 1). https://doi.org/10.1016/j.annemergmed.2016.05.010

Feng, L., Yuan, J., & Yang, L. (2020). An observation framework for retracted publications in multiple dimensions. *Scientometrics*, *125*(2). https://doi.org/10.1007/s11192-020-03702-3

Geng, Y., Cao, R., Han, X., Tian, W., Zhang, G., & Wang, X. (2022). Scientists are working overtime: when do scientists download scientific papers? *Scientometrics*, *127*(11). https://doi.org/10.1007/s11192-022-04524-1

Greshake, B. (2017). Looking into Pandora's Box: The Content of Sci-Hub and its Usage. *F1000Research*, *6*. https://doi.org/10.12688/f1000research.11366.1

Himmelstein, D. S., Romero, A. R., Levernier, J. G., Munro, T. A., McLaughlin, S. R., Greshake Tzovaras, B., & Greene, C. S. (2018). Sci-Hub provides access to nearly all scholarly literature. *ELife*, *7*. https://doi.org/10.7554/eLife.32822

King, E. G., Oransky, I., Sachs, T. E., Farber, A., Flynn, D. B., Abritis, A., Kalish, J. A., & Siracuse, J. J. (2018). Analysis of retracted articles in the surgical literature. *The American Journal of Surgery*, *216*(5), 851–855. https://doi.org/10.1016/j.amjsurg.2017.11.033

Kirsop, B., & Chan, L. (2005). Transforming access to research literature for developing countries. *Serials Review*, *31*(4). https://doi.org/10.1016/j.serrev.2005.09.003

Kleinert, S. (2009). COPE's retraction guidelines. *The Lancet*, *374*(9705), 1876–1877. https://doi.org/10.1016/S0140-6736(09)62074-2

Kohl, C. B. S., & Faggion, C. M. (2024). A comprehensive overview of studies that assessed article retractions within the biomedical sciences. *Accountability in Research*, *31*(6). https://doi.org/10.1080/08989621.2022.2154660

Mejia, C. R., Valladares-Garrido, M. J., Miñan-Tapia, A., Serrano, F. T., Tobler-Gómez, L. E., Pereda-Castro, W., Mendoza-Flores, C. R., Mundaca-Manay, M. Y., & Valladares-Garrido, D. (2017). Use, knowledge, and perception of the scientific contribution of Sci-Hub in medical students: Study in six countries in Latin America. *PLoS ONE*, *12*(10). https://doi.org/10.1371/journal.pone.0185673

National Information Standards Organization. (2024, June 27). *NISO RP-45-2024, Communication of Retractions, Removals, and Expressions of Concern (CREC): A Recommended Practice of the National Information Standards Organization*. https://doi.org/10.3789/niso-rp-45-2024

Owens, B. (2022). Sci-Hub downloads show countries where pirate paper site is most used. *Nature*. https://doi.org/10.1038/d41586-022-00556-y

Rosenkrantz, A. B. (2016). Retracted publications within radiology journals. *American Journal of Roentgenology*, *206*(2). https://doi.org/10.2214/AJR.15.15163

Sagemüller, F., Meißner, L., & Mußhoff, O. (2021). Where Can the Crow Make Friends? Sci-Hub's Activities in the Library of Development Studies and its Implications for the Field. *Development and Change*, *52*(3). https://doi.org/10.1111/dech.12638

Steen, R. G. (2011). Retractions in the medical literature: How many patients are put at risk by flawed research? *Journal of Medical Ethics*, *37*(11). https://doi.org/10.1136/jme.2011.043133

Suelzer, E. M., Deal, J., Hanus, K., Ruggeri, B. E., & Witkowski, E. (2021). Challenges in Identifying the Retracted Status of an Article. *JAMA Network Open*, *4*(6). https://doi.org/10.1001/jamanetworkopen.2021.15648

Taros, T., Zoppo, C., Yee, N., Hanna, J., & MacGinnis, C. (2023). Retracted Covid-19 articles: significantly more cited than other articles within their journal of origin. *Scientometrics*, *128*(5). https://doi.org/10.1007/s11192-023-04707-4

Tripathi, M., Dwivedi, G., Sonkar, S. K., & Kumar, S. (2018). Analysing retraction notices of scholarly journals: A study. *DESIDOC Journal of Library and Information Technology*, *38*(5). https://doi.org/10.14429/djlit.38.5.13103

Tripathi, M., Sonkar, S. K., & Kumar, S. (2019). A cross sectional study of retraction notices of scholarly journals of science. *DESIDOC Journal of Library & Information Technology*, *39*(2), 74–81. https://doi.org/10.14429/djlit.39.2.14000

Wager, E., Barbour, V., Yentis, S., Kleinert, S., & Council, C. (2010). *Retractions : guidance from the Committee on Publication Ethics (COPE)*. 64–66. https://doi.org/10.1111/j.1467-789X.2009.00702.x

Wager, E., & Williams, P. (2011). Why and how do journals retract articles? An analysis of Medline retractions 1988-2008. *Journal of Medical Ethics*. https://doi.org/10.1136/jme.2010.040964

Wan, S. (2022). Which Nationals Use Sci-Hub Mostly? *Serials Librarian*, *83*(3–4). https://doi.org/10.1080/0361526X.2023.2193613

Xu, S., & Hu, G. (2022). A cross-disciplinary and severity-based study of author-related reasons for retraction. *Accountability in Research*, *29*(8). https://doi.org/10.1080/08989621.2021.1952870

Figure 1: Extent of availability of different versions of articles in Sci-Hub
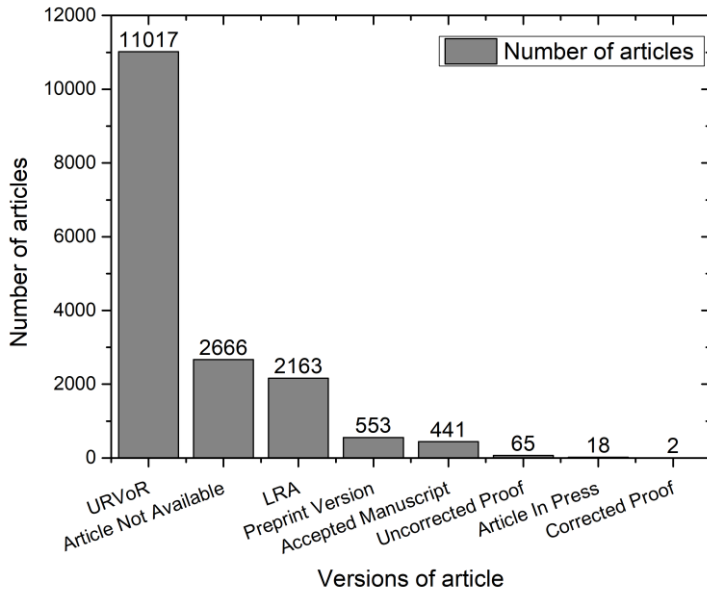
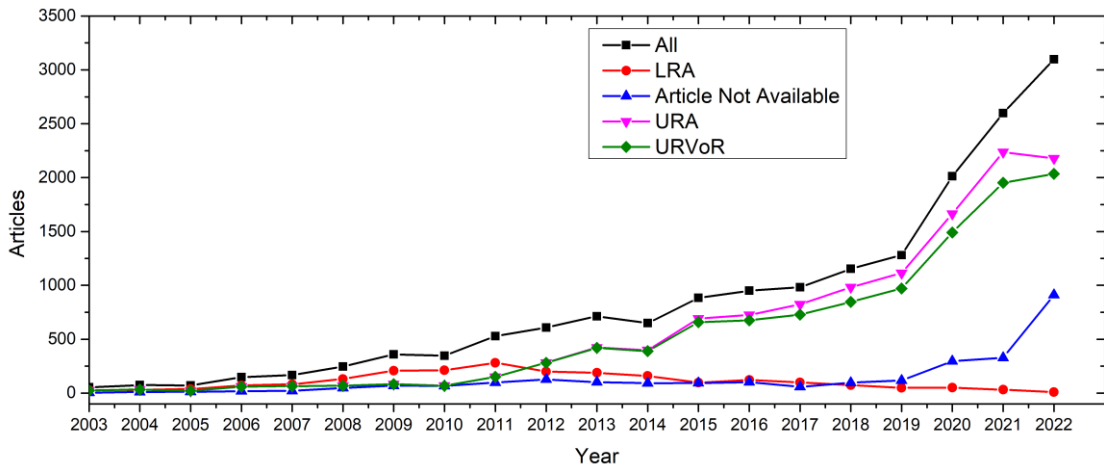Figure 2: Chronological growth of different versions of retracted articles

Figure 3: Subject-wise distribution of URA
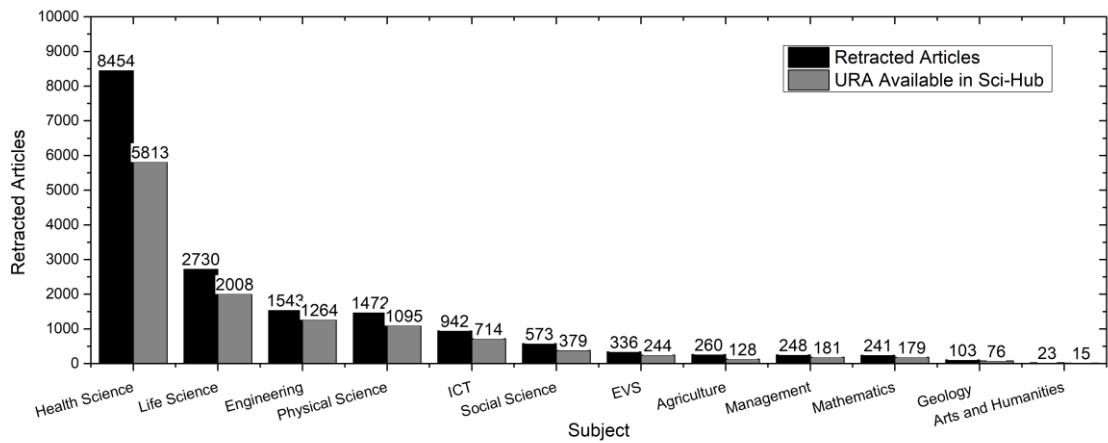
Table 1: Number of articles available on Sci-Hub as per its website on 10th April 2024.

Subject	No of articles
Medicine	24557530
Chemistry	16460921
Biology	15499507
Humanities	12592316
Physics	8658518
Engineering	6892853
Mathematics	3676789
Ecology	3676789
Computer Science	2768241
Economics	2572842
Geoscience	2571177

Table 2; Extent of availability of different versions of articles in Sci-Hub; URA=Sl. Nos (1+4+5+6+7+8)

Sl. No	Versions of Articles	Number of Articles
1	URVoR	11017 (65.09%)
2	Article not available	2666 (15.75%)
3	LRA	2163 (12.78%)
4	Preprint Version	553 (3.27%)
5	Accepted Manuscript	441 (2.61%)
6	Uncorrected Proof	65 (0.38%)
7	Article in Press	18 (0.11%)
8	Corrected Proof	2 (0.01%)
	Total	**16925**

Table 3: Year-wise distribution of different versions of retracted articles

Year	LRA (a)	Article Not Available (b)	URVoR (c)	Acc. Manuscript (d)	Article in Press (e)	Corrected Proof (f)	Preprint (g)	Uncorrected Proof (h)	URA (i)=(c+d+e+f+g+h)	All (a+b+i)
			Other Versions (URA - URVoR)							
2003	27	3	23	0	0	0	1	0	24	54
2004	30	11	34	0	0	0	0	0	34	75
2005	38	12	21	0	0	0	0	0	21	71
2006	71	18	59	0	0	0	1	0	60	149
2007	82	20	65	0	0	0	0	0	65	167
2008	131	47	68	0	0	0	0	0	68	246
2009	208	69	83	0	0	0	0	0	83	360

2010	212	66	67	0	0	0	1	0	68	346
2011	281	97	151	0	0	0	0	0	151	529
2012	200	127	281	1	0	0	0	0	282	609
2013	188	102	419	1	0	0	3	0	423	713
2014	159	92	389	1	1	0	6	1	398	649
2015	98	93	658	5	1	0	28	0	692	883
2016	122	102	675	42	2	0	7	0	726	950
2017	100	58	728	81	0	0	15	0	824	982
2018	75	96	846	71	4	0	58	3	982	1153
2019	49	117	971	43	10	0	90	0	1114	1280
2020	51	297	1492	80	0	0	91	2	1665	2013
2021	32	328	1952	106	0	2	129	49	2238	2598
2022	9	911	2035	10	0	0	123	10	2178	3098
Total	**2163**	**2666**	**11017**	**441**	**18**	**2**	**553**	**65**	**12096**	**16925**

Table 4: Subject-wise distribution of retracted articles

Subject of the article	LRA (a)	Article Not Available (b)	URA (c)	RA (a+b+c)	Percentage of RA in the subject out of total RA identified (RA/16925*100)	Percentage of URA against RA identified with in the subject (URA/RA*100)
Agriculture	24	108	128	260	1.54	49.23
Arts and Humanities	3	5	15	23	0.13	65.22
Engineering	144	135	1264	1543	9.12	81.92
EVS	36	56	244	336	1.98	72.62
Geology	12	15	76	103	0.61	73.79
Health Science	1087	1554	5813	8454	49.95	68.76
ICT	47	181	714	942	5.56	75.8
Life Science	418	304	2008	2730	16.13	73.55
Management	36	31	181	248	1.47	72.98
Mathematics	32	30	179	241	1.42	74.27
Physical Science	227	150	1095	1472	8.7	74.39
Social Science	97	97	379	573	3.39	66.14
Total	**2163**	**2666**	**12096**	**16925**	**100**	**71.47**

Table 5: Distribution of retracted articles in health sciences

Year	LRA (a)	Article Not Available (b)	URVoR (c)	Accepted Manu	Article In	Corrected Proof (f)	Preprint (g)	Uncorrected Proof (h)	URA (i) =(c+d+e+f+g+h)	All (a+b+i)

				script (d)	Press (e)					
				Other Categories (URA – URVoR)						
2003	10	2	6	0	0	0	0	0	6	18
2004	14	6	21	0	0	0	0	0	21	41
2005	19	10	12	0	0	0	0	0	12	41
2006	34	11	27	0	0	0	0	0	27	72
2007	32	14	32	0	0	0	0	0	32	78
2008	64	32	47	0	0	0	0	0	47	143
2009	109	47	46	0	0	0	0	0	46	202
2010	100	35	37	0	0	0	0	0	37	172
2011	159	62	78	0	0	0	0	0	78	299
2012	96	64	141	0	0	0	0	0	141	301
2013	111	49	208	1	0	0	0	0	209	369
2014	73	52	141	1	0	0	4	1	147	272
2015	51	42	301	1	0	0	7	0	309	402
2016	65	68	341	11	1	0	5	0	358	491
2017	57	37	369	29	0	0	7	0	405	499
2018	39	55	347	18	4	0	22	3	394	488
2019	19	65	424	12	7	0	31	0	474	558
2020	21	248	729	32	0	0	43	0	804	1073
2021	13	231	1031	38	0	0	87	25	1181	1425
2022	1	424	1004	8	0	0	65	8	1085	1510
Total	**1087**	**1554**	**5342**	**151**	**12**	**0**	**271**	**37**	**5813**	**8454**

Table 7; Extent of retracted articles (RA) and URA of top 10 publishers from 2003 to 2023

Year	Springer - Nature		Elsevier		Wiley		T&F		Sage		PLOS		WK		OUP		Dove		Spandidos	
	RA	URA	RA	URA	RA	URA	RA	URA	RA	URA	RA	URA	RA	URA	RA	URA	RA	URA	RA	URA
2003	13	3	13	8	6	5	1	0	0	0	0	0	1	0	0	0	0	0	0	0
2004	8	6	25	12	8	5	2	1	1	1	0	0	3	3	7	3	0	0	0	0
2005	7	2	21	2	11	7	1	0	0	0	0	0	4	3	3	0	0	0	0	0
2006	16	8	21	4	24	10	6	4	1	1	0	0	6	3	7	4	0	0	0	0
2007	16	13	54	7	26	14	5	2	2	2	2	1	2	1	6	4	0	0	0	0
2008	23	12	82	12	43	11	11	5	3	2	0	0	14	4	15	1	0	0	0	0

2009	56	18	80	9	60	21	38	14	2	0	1	1	26	2	16	2	0	0	0	0
2010	55	11	88	9	44	9	14	4	6	0	3	3	19	7	10	3	0	0	0	0
2011	76	21	130	27	76	27	33	11	15	4	3	3	47	11	22	7	0	0	0	0
2012	125	81	106	44	85	73	65	12	21	5	14	11	19	10	19	8	1	0	2	1
2013	155	114	209	154	70	55	49	18	11	2	9	9	43	9	18	11	1	0	1	0
2014	120	101	128	114	59	45	40	16	70	34	12	3	24	8	20	11	3	0	3	1
2015	281	247	187	173	86	70	29	23	35	31	8	5	26	16	18	8	4	2	7	3
2016	227	205	182	175	105	68	62	54	20	15	18	3	19	11	19	19	9	0	12	9
2017	263	256	206	195	114	89	56	49	24	20	20	4	25	16	19	15	11	1	10	8
2018	158	143	318	312	117	105	48	39	38	35	52	31	34	20	26	24	6	1	16	14
2019	234	220	266	236	133	125	65	64	52	50	105	103	39	24	40	34	12	10	11	8
2020	352	336	364	346	208	186	133	129	64	59	84	82	63	45	47	43	73	68	34	25
2021	424	407	317	295	398	385	141	138	198	193	45	40	54	27	34	28	110	90	130	121
2022	925	815	466	393	502	197	169	145	43	26	213	92	34	16	28	23	139	110	101	89
Total	3534	3019	3263	2527	2175	1507	968	728	606	480	589	391	502	236	374	248	369	282	327	279