Annals of Library and Information Studies
Vol. 69, December 2022, pp. 304-322
DOI: 10.56042/alis.v69i4.67478

Retractions in India since independence: a multifaceted analysis for 75 years through data carpentry

Parthasarathi Mukhopadhyay[a], Mondrita Mukhopadhyay[b] and Mustak Ahmed[c]

[a]Professor, Department of Library and Information Science, Kalyani University, West Bengal, India, Email: psm@klyuniv.ac.in

[b]University Research Scholar, Department of Library and Information Science, Kalyani University, West Bengal, India, Email mondrita24c@gmail.com

[c]Senior Research Fellow, Department of Library and Information Science, Kalyani University, West Bengal, India, Email: mustak.masu@gmail.com

Received: 14 October 2022; revised & accepted:31 December 2022

This study aims to explore the nature of article retractions in India for a time frame of 75 years (1947–2021) by developing a comprehensive primary dataset of 1,376 retracted items, and then merging the dataset as developed with an array of external datasets by deploying data carpentry methods and techniques available in OpenRefine, an open-source data wrangling software. This value addition leads to exploring many new angles of study related to retraction in India, like gender distribution, geospatial distribution, institutional distribution, subject distribution in retraction, relations between journal quality and retraction, identification of serial offenders, relations between citation and retraction, and more importantly, reasons for retraction. It discovers many facts about retraction in India and attempts to represent the findings using a few next-generation visualization techniques. The major findings include the following: retraction in India is growing exponentially, and we are now the 4th highest in retraction on a global scale; the majority of the retracted items are published in quality journals in terms of quartile, impact factor, H-Index, and CiteScore; retracted items are distributed in both close-access and open-access source titles; most retracted items are able to fetch citations (mean citation 19.73), including recent citations; a considerable number of retracted items are authored by serial offenders; the retraction map of India includes the majority of the states and union territories; elite educational and research institutes are equally responsible for retraction, along with not-so-known institutes; text manipulation is still the most visible reason for retraction in India, but data manipulation, and image manipulation are increasing rapidly. It also finds that a few cases of retraction are simply due to a lack of awareness on the part of the scholars.

Keywords: Retraction; Research Ethics; Data carpentry

Introduction

Studies on different aspects of retraction of scholarly contributions are increasing all over the world owing to the advent of the database from Retraction Watch (http://retractiondatabase.org) and tools like scite (https://scite.ai/), re-cite (https://github.com/recite/re-cite.org), the preprint health check facility in Scholarcy (https://app.scholarcy.com/preprint-check.html) and Problematic Paper Screener (https://dbrech.irit.fr/pls/apex/f?p=9999:1:::NO:::). The Pubpeer (pubpeer.com) website serves as an open forum for reporting academic misconduct in relation to a published paper. It allows searching for a given paper by DOI, PubMed ID, or ar Xiv ID and also provides a browser extension to track retracted papers. The integration of retraction databases with browser extensions like LibKey Nomad, Lean Library, CrossMark and reference management tools like Zotero (zotero.org) is making it increasingly easier to detect and avoid retracted papers.

Apart from the above tools and datasets, three API-based accesses to retraction status require special mention in the context of this study – CrossRef API[1], PubMed API[2] and Open retraction API[3]. A data-intensive and comprehensive study of retraction in India is necessary not only to have a panoramic view but also for developing policies related to academic integrity. For example, the academic integrity policies of UGC[4] and AICTE[5] talk only about plagiarism, whereas this study proves that plagiarism is not the only major concern related to retraction of the papers published in India.

As of September 30th 2022, the retraction watch database has reported a total of 1,541 retracted items originated from India during the period from August 15, 1947 to August 14, 2021 (75 years), which is

much lower in comparison with China (14,728 items) and the United States (4,908 items), but higher than the UK (1,358 items), Japan (1,303 items), Germany (966 items) and South Korea (717 items) in the same time-frame. According to the retraction watch database[6], India is now in the 4th highest position in terms of retraction, after China, US and Russia.

For the period August 15th 2021 to August 14th 2022 (immediate next year to the period of this study), RetractionWatch database shows that a total of 913 items have been retracted in this time period, and 350 papers out of these 913 items were retracted on a single day (February 23rd 2022) by the publisher Institute of Physics (IOP) from two different 2021 conference proceedings—*Journal of Physics: Conference Series* (232 contributions), and *IOP Conference Series: Materials Science and Engineering* (118 contributions)[7]. All of these contributions have been retracted for a new genre of academic misconduct called *tortured phrases* conducted with the help of a set of computer programs ("tortured phrases are what happens to words that get translated from English into a foreign language, then back to English"[8]).

Review of literature

The studies related to identifying patterns in retraction require comprehensive datasets, and many such studies have been conducted in recent times. For example, a recent study of Vuong et al[9] complied a global dataset of 18,603 retracted items between the years 1753 and 2019 covering 127 research fields and found that the rise of retraction started in 1999 and an unusual number of retractions happened in 2010. According to this study, almost 60% of all retracted papers were from publishers like the Institute of Electrical and Electronics Engineers (IEEE), Elsevier, and Springer, with the highest number of retractions from the publications of IEEE, and "fake peer-review" was the major reason for these retractions.

Another large-scale study that dealt with the retracted papers in scientific publications between 2001 and 2010 issued an early alert and found that retractions were increasing faster than the number of global scientific publications, and in three countries, namely China, India, and South Korea, retractions were higher than the global average[10]. A study in 2014 based on ScienceDirect database found 995 numbers of retracted papers with three major reasons for retraction - ethical misconduct (64%), scientific distortion (31%), and administrative issues (5%)[11]. A series of papers using different datasets in different time frames were attempted to explore the reasons for retractions and it was found that the major reasons were academic misconduct, unintentional errors, unreliable data (data falsification and fabrication) and plagiarism[12–15].

However, a study[16] that attempted to categorize reasons for retractions across the disciplines divided the possible reasons into four categories—papers with misconduct and scientific error (category I), misconduct and no scientific error (category II), no misconduct and no scientific error (category III), and no misconduct and scientific error (category IV), and found that the maximum had scientific errors (category I & IV, 93.62%), misconduct accounted for almost half of the cases (category I & II, 44.2%), and only 32% had neither scientific error nor misconduct. There are some interesting findings related to retraction like the majority of retracted papers have multiple authors[17]; China has the fastest growing retraction rate in comparison to other countries[17]; co-authors of retracted papers suffer academically due to misconduct by their colleagues[18]; larger author groups have a lower retraction probability in comparison to smaller author groups[19]; one-fourth of retracted papers cite other retracted papers[20]; and scientific misconduct happened in those countries that had lack of research integrity policies and countries where individual publication performance was rewarded with cash[21].

Many studies have been conducted to understand why retracted papers get citations. For example, a study of biomedical literature from PubMed has revealed that almost 16% of the citations received by the retracted papers in biomedical discipline are substantive citations (cited in methodology, results, or in other important sections) and 80% of the citations are tacit in nature (only part of a literature review), but interestingly, 19.67 was the mean citations per retracted article as per this study[22]. Another study pointed out that citations for retracted papers decreased post-retraction, and most of the citations came from other countries, which were different from publication countries[23].

Most of the researchers[24–26] who studied post-retraction citations agreed that delays in retraction are responsible for citations in pre-retraction period, and a lack of proper notification for retraction is the most visible reason for receiving citations even after the

paper has been retracted. A recent study[27] reported that the process of retraction is not quite organized yet and more than half of the retracted Covid-19 research articles are available in full-text without proper retraction notices.

Many researchers attempted to explore global and country-level scenarios in retraction such as a study of retraction watch database (2013-2015) found that retracted papers belonged to 71 countries and 15 countries had the maximum number of retractions[28]; an analysis of retraction in life sciences identified that China, Japan, India, and Germany occupy the first five ranks[29]; another study that dealt with the papers retracted due to plagiarism and duplicate publication across the nations (53 countries) found that India is now the 6th highest contributor[30]. There is, however, only one study[31] available that deals with the reaction profile of India exclusively. The authors of this study collected a set of 239 retracted items (included 195 journal papers and 44 conference papers retracted between 2005 to 2018 from Scopus database) and found that retractions have increased manifold in India since 2010, most of the retracted papers had multiple authors from different countries, and plagiarism was the most visible reason for retraction along with image manipulation.

Objectives of the study

The broad objective of this study, as specified in the title, is to discover patterns of retraction in India from different points of view, which so far have not been explored by any research studies on retraction in India. It aims to develop a comprehensive primary dataset of items published in India and subsequently retracted for the period from August 15th 1947 to August 14th 2021 (75 years of independence). The multifaceted analyses are ensured through the merging of the primary dataset with an array of data sources like citations, altmetric, open-access status, rich external metadata sets, global-scale journal indexes, name-to-gender and geospatial datasets. The processes of data fetching and data merging are based on data carpentry methods (data gathering, data curation, data faceting, etc.) and data carpentry tools (OpenRefine, Tableau, Python plotting and graph libraries). However, the specific objectives of this research study are listed here under three broad groups:

1. Gr. A: To develop an all-inclusive primary dataset of retracted items that originated in India since its independence up to the completion of 75 years (August 14th 2021) by consulting an array of sources and then confirming the retraction status from the respective source titles;
2. Gr. B: To enrich the primary dataset through REST/API-based content negotiation from a set of selected external data sources in order to discover different visible and hidden patterns of retraction in India (for example, citations profile, journal quartiles, authorship in retraction, genders in retraction, serial offenders, and many more); and
3. Gr. C: To record, represent, and visualize trends, patterns, and growth in retraction in India over the last 75 years, including an analysis of reasons for retraction and distribution of retraction subject-wise, institutionally, and geographically.

Methods

This research study is based on the prime principle of data wrangling: *merging related data from multiple sources enhances the value of a target dataset*. We divided the methods used into three groups of activities: a) the development of a comprehensive primary dataset of retracted papers for the period 1947-2021 (=75 years) published in India; b) the integration of the primary database with many other ODbL-based datasets, like citations (dimensions.ai, scite.ai & open citation corpus), altmetric attention scores (altmetric.com), open-access status (unpaywall.org), metadata (crossref.org) gender status (genderize.io), geospatial datasets (nominatim.openstreetmap.org), journal quartile (scimagojr.com), global journal indexes (Scopus and DOAJ) etc., for enabling multifaceted analysis; and c) the analysis and visualization of the consolidated retraction dataset to discover the patterns of retraction in India.

Group A: Developing the primary dataset

We retrieved primary retraction data from the Retraction Watch database[6], confirmed retraction status from a variety of datasets such as CrossMark, PubPeer, and Open Retraction, and finally obtained the retraction notice from the respective journal portals. The Retraction Watch database[6] allows filtering of results by country of publication and range of retracted years but does not allow data download or data scraping. Moreover, it retrieves only 50 results at a time for a query. Therefore, we prepared a CSV file on retracted papers semi-automatically in consultation

with the above-mentioned data sources. Sometimes, the Retraction Watch database includes more than one mention of a retracted item, as it enters retracted items by their related DOIs, and a retracted paper may have the original DOI, the DOI of the retraction notice, and the DOI of the authors' response.

The primary dataset thus developed (as a csv file) was imported into OpenRefine and curated through a collision-detection algorithm to include a retracted paper only once in the dataset. A total of **1,541** retracted papers that originated in India and were retracted from August 15th 1947 to August 14th 2021 (75 years) were collected initially from the Retraction Watch database (as of 30th September 2022), and after duplicate checking by using DOI and PMID as unique keys and confirmation of retraction from respective journal portals, it finally settled down to 1,376 retracted items.

The primary datasets of 1,376 retracted papers include the following items of information (data elements) after processing and curation at OpenRefine's end: i) Item title, ii) Author(s), iii) Affiliated institution(s) with detail address, iv) Journal name, v) Publisher name, vi) Reason(s) for retraction, vii) Article type, viii) Nature of retraction notice, ix) Country/Countries of the author(s), x) Publication date (in ISO format), xi) Original DOI/PMID, xii) Subject groups, xiii) Retraction date (in ISO format), xiv) Retraction DOI/PMID, and xv) Local ID (arranged locally to identify each record of retraction uniquely, in view of the need for such an ID for developing the secondary datasets). A few interesting observations as experienced by the authors during the development of the primary dataset may be mentioned here:

I) The dataset of 1,376 retracted papers includes 1,267 items having DOIs (92.07% of the final primary dataset), 749 items having PMID (54.43%), 715 items having both a DOI and a PMID (51.96%), 34 items having either a DOI or a PMID (2.47%), and 75 items not having any DOI or PMID (5.45%). This information is critical because the next level of work necessitates at least one unique ID of a paper for data wrangling activities.

II) Open Retraction (openretractions.com), which is based on the CrossRef and PubMed databases, provides REST/API-based responses against a syntax like *"http://openretractions.com/api/doi/" + value +"/data.json"* (value is the DOI or PMID of the paper) and produces responses in JSON format with retraction status in the title like *{"title":"RETRACTED: Analysis of codeposited Gd2O3/SiO2 composite thin films by phase modulated spectroscopic ellipsometric technique","doi": "10.1016/j.apsusc.2006.03.013" …. }*. There is no separate data element to indicate the retraction status of a published item.

III) Open Retraction includes records of 1,14,596 retracted papers (as of September 30th 2022), but is not quite comprehensive as far as retracted records from India are concerned; for example, REST/API-based data fetching from Open Retraction for 1,301 items (1,267 + 34) with DOI or PMID provided results for only 381 items (29.28%);

IV) The CrossRef metadata response (possibly the most comprehensive metadata sets for published papers with DOIs) does not include any exclusive data element for retraction status but rather includes it in the *title data field* inconsistently. A query to the CrossRef forum reveals that they use the metadata supplied by the publishers on an as-is basis, and indication of retraction status is optional, though CrossRef requests all participating publishers to follow the COPE guidelines for retraction[32].

It has been decided at this point that the generic analysis of retracted papers like geographical distributions, authorship patterns, inter-country collaborations, reasons for retractions, retraction delay analysis, journal-wise and publisher-wise distributions, gender analysis, etc. will include all 1,376 retracted papers, but the value additions like subject-wise analysis, citation and altmetric data analysis, open access status etc. will be based on the 1,267 retracted papers having DOIs (as DOI is the only key for farther data wrangling activities).

Group B: Developing the secondary datasets

As the primary dataset alone is inadequate to achieve the objectives of the study, value was added by including related items of information from other sources. The ODbL-based data sources, syntax for content negotiation, target data elements and purpose of these datasets for 1,267 retracted items with DOI are represented in Table 1.

Apart from these ten ODbL-based data sources, the method includes integration of three more data sources, namely public data dump of Scopus source titles, DOAJ-listed journal metadata and Scimago dataset. All these datasets are available as CSV downloads and merged with the primary dataset based

Table 1 — Content negotiation for ODbL-based based data sources

Sl. no.	Data source	REST/API syntax in OpenRefine	Purpose	Target data elements	No. of responses
1	Open Retraction	"http://openretractions.com/api/doi/" + <value> +"/data.json"	Retraction status	1. Retraction status 2. Retraction type	381 (against 1,267 queries)
2	Unpaywall	"https://api.unpaywall.org/v2/" + <value> + "?email=<your mail-id-here>"	Open access (OA) status	1. OA status 2. OA category 3. OA repository data 4. OA license	1.242 (against 1,267 queries)
3	Scite	"https://api.scite.ai/tallies/" + <value>	Citation data	1. Total citations 2. Supporting citations 3. Contradicting citations 4. Mentioning citations	1.242 (against 1,267 queries)
4	Dimensions	"https://metrics-api.dimensions.ai/doi/" + <value>	Citation data	1. Total citations 2. Recent citations	1.248 (against 1,267 queries)
5	Open Citation Corpus	"https://opencitations.net/index/api/v1/citation-count/" + <value>	Citation data	1. Total citations	1.262 (against 1,267 queries)
6	Altmetric	"https://api.altmetric.com/v1/doi/" + <value>	Altmetric data	1. Altmetric attention score 2. Subjects of the items	589 (against 1,267 queries)
7	CrossRef	"https://api.crossref.org/works/" + <value>	Item metadata	1. Retraction status (in title, or abstract) 2. Subject keywords	1.240 (against 1,267 queries)
8	Semantic Scholar	"https://api.semanticscholar.org/v1/paper/" + <value>	Item metadata	1. Retraction status 2. Subject keywords 3. Fields of study	896 (against 1,267 queries)
9	Nominatim	"https://nominatim.openstreetmap.org/search.php?q=" + value + "&format=jsonv2&limit=1&country=India"	Geospatial data for map based visualization	1. Display name of place 2. Bounding box data 3. Longitude 4. Latitude	2,372 (against 2,372 places)
10	Genderize	"https://api.genderize.io/?name=" + value.escape('url') + "&country_id=IN"	Name-to-gender inference	1. Gender 2. Probability of correctness	5,292 (against 5,599 author names)

Note: For Sl. no. 1 to 8 <value> is DOI; for SL no. 9 <value> is place/city name; for Sl no. 10 <value> is author name.

on a composite primary key combining source title and ISSN/ISBN. The merging was accomplished by *Cross function* feature of OpenRefine to add selected data elements into the primary dataset from other projects of OpenRefine on the basis of a unique key.

Results

The results as obtained through the steps of data wrangling, data curation, data extraction, data faceting and data processing are represented here under 11 interrelated heads.

Growth of retraction

An analysis of the retracted papers in the primary dataset of 1,376 items in the year range 1947-2021 shows that it is growing exponentially (Fig. 1A), conforming to the results of studies that reported the alarming rates of retraction in some countries, including India[10,17,28-31]. Fig. 1B shows a comparison of the retraction growth rate in the world and in India. The first incident of retraction in India occurred as early as in 1992 (1 document) and reached its peak in 2014 with 177 retractions (retraction years). These 1,376 items were published in the year range January 1, 1989, to June 02, 2021, with the highest number of papers retracted in 2014 (132 items), followed by 2012 (131 items), 2013 (122 items), 2015 and 2018 (115 items published in each year).

Delays in retraction

Retraction is a complex process. It involves reporting of misconduct or errors, authors' responses, editors' communication, decision making, and notification, and often takes a considerable amount of time. As this dataset includes the date of publication and the date of retraction for each record of retraction, it is possible to throw light on retraction delays to get an idea about the entire process. The data wrangling software OpenRefine allows for the conversion of a date into ISO format under the *common transforms*

Table 2 — Calculation of retraction delay

Date of publication (Pub_Date)	Date of Retraction (Ret_Date)	GREL	Result
1996-11-01T00:00:00Z	1997-06-10T00:00:00Z	diff(cells["Ret_Date"].value,cells["Pub_Date"].value, "**months**")	7
		diff(cells["Ret_Date"].value,cells["Pub_Date"].value, "**days**")	221

Table 3 — Retraction delay in days

Range in days	No. of retracted items
0-100	216
101-500	453
501-1000	252
1001-1500	159
1501-2000	99
2001-2500	66
2501-3000	51
3001-8133	80
Total	1,376

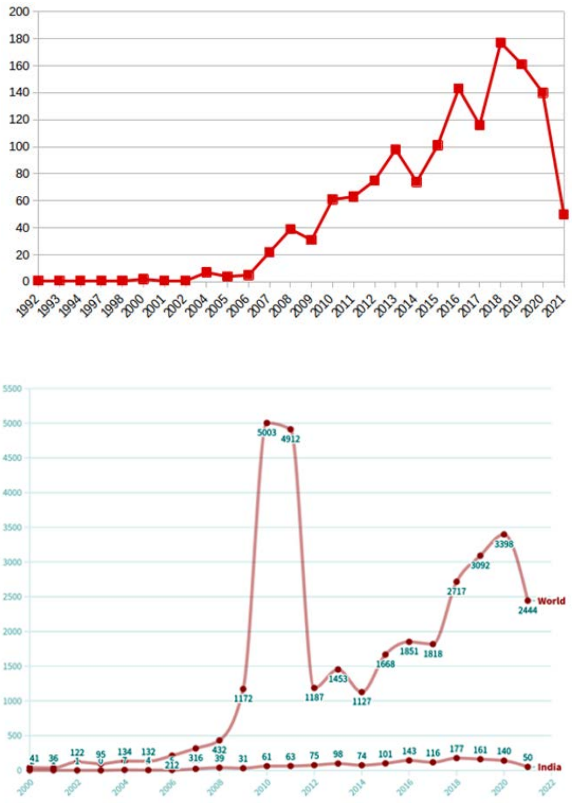

Fig. 1 (a-b) — Exponential growth of retraction in India (January 1992 - to Aug. 14th 2021); Growth in retraction – World vs. India (January 2000 - to Aug. 14th 2021)

facilities, and simple GRELs can then be used to calculate the difference in days, months, or years (Table 2).

The highest delay is 8133 days (1 paper), whereas 51 papers have been retracted on the same day. An analysis (Table 3 and Figure 2) shows that only

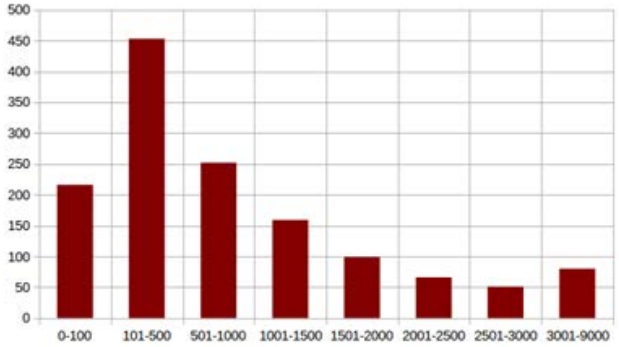

Fig. 2—Retraction delay in days (day ranges)

15.69% of the items (216 items) are retracted within the first 100 days; the majority of the items were retracted within the range of 101-500 days (32.92%, 453 items), whereas 5.81% of the items (80 items) took a bit longer, of more than 3000 days.

Authorship patterns in retraction

A total of 5,599 instances of authorship *are associated* with the 1,376 retracted items that were published from India during the period 1989-2021 and have been retracted during the period from June 1992 to August 14, 2021. It means that most retracted papers from India have multiple authors, with an average of slightly more than four (4.06 to be exact) authors per paper. There is 1 paper with 32 *authors, whereas* 110 papers are single-authored publications (7.99%); 279 papers are contributed by 2 authors (20.27%); 317 papers (23.03%, the highest in the dataset) are partnered by 3 authors; and 237 papers (17.22%) are produced through 4-author collaboration. A detail view of 20 instances of authorship in the dataset is given in Table 4 and Figure 3.

Almost 80% of the total retracted papers (79.72% to be exact) have authorship by 2, 3, 4 and 5 authors, covering almost 60% of authorship (3,337 authorship, 59.59% of total authorship), and the rest 20% of the retracted papers (20.28% to be exact) are conferred to 2,262 authorship (40.40% of total authorship). A significant point to be noted here is that these 5,599 instances of authorship include 4,304 unique authors as some authors have repeat retractions, like one author in this dataset, who alone, has 28 retracted papers.

Table 4 — Authorships in retraction

Sl. nos.	No. of author(s)	No. of paper(s)	Total no. of authors	Sl. nos.	No. of author(s)	No. of paper(s)	Total no. of authors
1	1	110	110	11	11	9	99
2	2	279	558	12	12	8	96
3	3	317	951	13	13	3	39
4	4	237	948	14	14	1	14
5	5	154	770	15	15	2	30
6	6	96	576	16	16	3	48
7	7	61	427	17	18	1	18
8	8	46	368	18	21	4	84
9	9	22	198	19	23	1	23
10	10	21	210	20	32	1	32
		Grand total				1,376	5,599

Table 5 — Repeat retractions

No. of instances of retraction	No. of associated authors	No. of instances of retraction	No. of associated authors	No. of instances of retraction	No. of associated authors
1	3641	7	12	15	1
2	421	8	8	19	2
3	78	9	1	23	1
4	63	10	2	26	1
5	26	11	2	28	1
6	21	12	2		

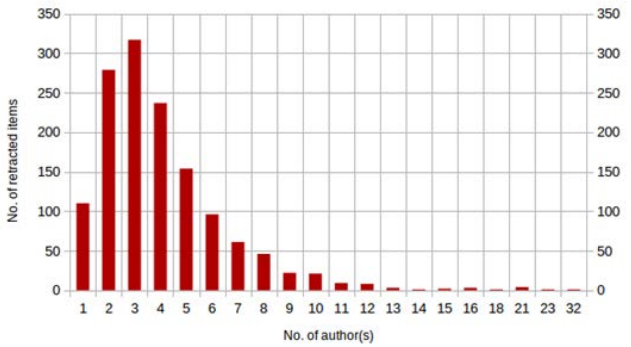

Fig. 3 — Authorship distribution in retraction

Repeatedly retracted authors

The reasons for retractions can be grouped into two basic categories: intentional academic misconduct of varying degrees and honest academic errors. Some authors are responsible for repeated acts of publication with deliberate academic fraud of varying degrees. According to Steen, these acts are *"neither naive, feckless nor inadvertent"* but *"a deliberate effort to deceive, a motivation fundamentally different from papers retracted for error"*[33]. Though there is no universal definition of how many retracted papers in the credit line of an author makes her/him a repeat offender, the Retraction Watch blog has created a concept of a Leaderboard that at a global-scale lists authors with 25 or more retracted items (https://retraction watch.com/the-retraction-watch-leaderboard/).

The Leaderboard lists an author from India who has the highest number of 28 retracted items in this dataset. The primary dataset contains 5,599 authors with 4,304 unique authors, of which 4,283 authors belong to different institutes in India. Fortunately, most of the Indian authors are so far associated with only one instance of retraction (3,641 authors, 85.01% of 4,283 Indian authors) and may be considered under the academic error category, but 642 authors (14.99%) have committed it more than once and therefore belong to the category of academic misconduct (Table 5).

If we place the threshold value at 5 or more instances of retraction by an individual author to be termed as repeat retraction, then there are 80 authors from 30 different institutes in India who are responsible for 49 retracted items (3.56% of 1,376 items). The top ten of these institutes, from where repeated retractions have been committed during the period of study (1947–2021) are listed in Table 6, and Fig. 4 illustrates the radial view for all 30 institutes.

Genderizing retractions

This part of the study that aims to explore the gender analysis of authorship of retracted papers published in India is based on a data carpentry method known as name-to-gender inference. A set of such services are recommended by Santamaria & Mihaljevic

Table 6 — Affiliations of serial offenders (top ten institutions)

Affiliated Institute	Category of Institute	City	State	No. of repeat offenders
Indian Institute of Technology (ISM) Dhanbad	Indian Institute of Technology	Dhanbad	Jharkhand	9
S.V. University	UGC-University (Public)	Tirupati	Andhra Pradesh	8
University of Rajasthan	UGC-University (Public)	Jaipur	Rajasthan	8
National Institute of Cholera and Enteric Diseases (NICED, Kolkata)	ICMR-Medical Research	Kolkata	West Bengal	6
Annamalai University	UGC-University (Public)	Annamalai Nagar	Tamil Nadu	4
Indian Institute of Toxicology Research (CSIR-IITR)	CSIR-Research Institutes	Lucknow	Uttar Pradesh	4
Aligarh Muslim University (AMU)	UGC-University (Public)	Aligarh	Uttar Pradesh	3
Bhabha Atomic Research Centre (BARC)	DAE-Research Organization	Mumbai	Maharashtra	3
Centre for DNA Fingerprinting and Diagnostics	DBT-Biotechnology Research	Hyderabad	Telangana	3
Kalasalingam Academy of Research and Education (Kalasalingam University)	Private-University & Institute	Krishnankoil	Tamil Nadu	3

Table 7 — API call response format of genderize.io (source: Mukhopadhyay et al[35])

Call	Response
https://api.genderize.io/?name=Uma&country_id=IN	{"name":"Uma","gender":"female","probability":0.66,"count":276,"country_id":"IN"}
https://api.genderize.io/?name=Reema&country_id=IN	{"name":"Reema","gender":"female","probability":0.98,"count":114,"country_id":"IN"}
https://api.genderize.io/?name=Saraswati&country_id=IN	{"name":"Saraswati","gender":"female","probability":1,"count":9,"country_id":"IN"}
https://api.genderize.io/?name=Polumetla&country_id=IN	{"name":"Polumetla","gender":null,"probability":0,"count":0,"country_id":"IN"}

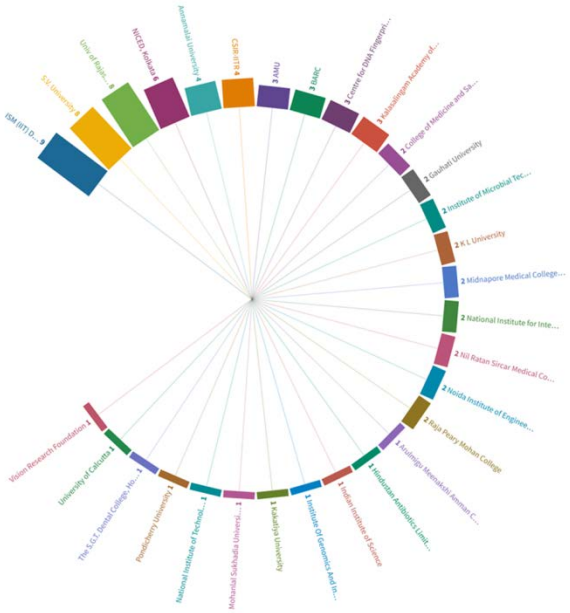

Fig. 4 — Institutional distribution of serial offenders

in their work[34] and Mukhopadhyay et al[35]. The list of recommendations in the said work includes five such services, namely Gender API (https://gender-api.com/), Gender-guesser (https://pypi.python.org/pypi/gender-guesser/), genderize.io (https://genderize.io/), NameAPI (https://www.nameapi.org/), and NamSor (http://www.namsor.com/), of which two inference services support API calls over GET requests (gender-api.com and genderize.io) and the other three services support POST requests. This study chose *genderize.io* from the GET group because it generously provides free API calls at a rate of 1000 per day, as opposed to the service gender-api.com's 500 free calls per month. Other key points in favour of *genderize.io* include: i) availability of the *probability* parameter (in the range of 0 to 1) within the response format; ii) the ability to limit queries to a specific country (by ISO 3166-1 country code); and iii) comparatively higher confidence in inferencing Asian names, including Southern Asia[34]. A set of typical API calls to *genderize.io* from the data wrangling software OpenRefine is given in Table 7 to clarify the probability parameter in the responses from *genderize.io* in JSON format.

The inference services, including *genderize.io*, generally determine gender based on the first name or given name of an author. The curated primary dataset comprises a total of 5,599 authors and includes

Table 8 — Gender analysis report for retracted papers from *genderize.io*

Condition	Result (N=5292)			
	Female	Male	Unsure	Female:Male
Extremely sure: if(value.parseJson().probability == 1,value.parseJson().gender, "Unsure")	406	1586	3300	1:3.90
Fairly sure: if(value.parseJson().probability >= 0.90,value.parseJson().gender, "Unsure")	938	3288	1066	1:3.50
Moderately sure: if(value.parseJson().probability >= 0.75,value.parseJson().gender, "Unsure")	1031	3430	831	1:3.32
Somehow sure: if(value.parseJson().probability >= 0.60,value.parseJson().gender, "Unsure")	1090	3495	707	1:3.20

Table 9 — Inter-country collaboration summary

India	+ 1 country	+ 2 countries	+ 3 countries	+ 4 countries	+ 6 countries	+ 7 countries	+ 8 countries
No. of retracted papers	171 (77.72%)	33 (15.00%)	8 (2.857%)	5 (3.63%)	1 (0.45%)	1 (0.45%)	1 (0.45%)
No. of countries involved	34	28	19	12	6	7	8

authors' names in three different styles (on the basis of the name elements as recorded in the published papers): a) first name/given name is somehow obvious e.g. Saraswati Ghosh or Satyanarayan Sankhwar (4,400 authors –78.58%); b) second part (e.g. N Ananthakrishnan or S Jagdish) or third part (P G Gopinath or G A Ravishankar) or fourth and subsequent parts (e.g. T N R Srinivas or R V M S S Kiran Kumar) of the name is actual name element (892 authors – 15.93%); and c) only family name (e.g. H K Verma or T K Chattopadhyaya) is available (307 authors - 5.48%).

Because the chosen name-to-gender inference service (*genderize.io*) depends on the first name or given name component of the personal name elements for gender analysis, a consolidated csv file has been developed as a subset of the master curated database which contains 5,292 (4,400 + 892) authors whose given name elements are available either from the first part or from subsequent name parts, and then a suitable GREL *"https://api.genderize.io/?name="* + *value* + *"&country_id=IN"* can fetch responses from the selected name-to-gender service into the data wrangling software. The responses available in JSON format are then analyzed to study female-male ratio in retraction based on the probability score (range is 0 to 1) and it has been decided that the gender report will only be considered if the probability is at least 0.60 at the lowest scale[35]. The result set is represented here under four categories (Table 8).

As expected, the settings of higher probability of correctness in condition are producing lesser number of definitive results and greater number of undetermined ("Unsure") reports for gender analysis. The mean value of the authorship gender data shows that the female-male ratio in Indian retraction is 1:3.40 (rounded to 2 decimal places).

Collaboration in retraction

A total of 1,156 retracted items (84.01%) out of 1,376 are exclusively of Indian origin, involving authors either from the same institute or from an array of institutes distributed throughout the country. The remaining 220 items of the dataset (15.98%) have collaborators from many different countries, ranging from 1 to 8 countries, involving 53 different countries. A summary table (Table 9) shows that most of the collaborative retracted papers are based on two-country partnerships (77.72%), and the three-country partnership is a distant second in the row (15.00%).

All these inter-country collaborations of 220 retracted items involve 53 distinct countries distributed across the globe (Figure 5), and analysis shows that the top ten collaborative countries in the context of retraction in India are – United States (69 items – 31.36% of collaborative retracted papers), Saudi Arabia (21 items – 9.54%), Malaysia and United Kingdom (15 items each – 6.81%), Australia, Germany and South Korea (14 items each – 6.36% each), Italy and Japan (10 items each – 4.54% each), and Iran (9 items – 4.09%). The other countries, very close to the top ten positions (in the context of collaboration in retraction with India) are: Taiwan (8 items – 3.63%); Brazil, Canada and China (7 items each – 3.18% each).

An in-depth analysis of two-country and three-

Fig. 5 — Inter-country collaboration in retraction (220 retracted items of collaboration)

country partnerships (together these two types of partnerships are responsible for almost 93% of the retracted papers that are produced in collaboration) shows in both the cases collaboration with United States leads the lists with 49 items in two-country collaboration and with 13 items for three-country collaboration, followed by Saudi Arabia with 14 items and United Kingdom with 7 items for two-country and three-country collaborations respectively. An analysis of the 1,156 retracted items of exclusive Indian origin, involving 867 institutions distributed throughout the country (a subset of the primary dataset of 1,376 items), shows that 648 items are produced by single institutions (56.05% of 1,156 items; with one or more authors from the same institution), whereas inter-institutional collaborations have produced 508 retracted items (43.94% of 1,156 items).

Source titles in retraction

The 1,376 retracted items under study were published in 816 source titles, including journals, conference proceedings, edited books, and so on. The source titles with 10 or more retracted papers of Indian origin are: The *Journal of Biological Chemistry* (39 items), *PLoS One* (33 items), *International Journal of Mechanical and Production Engineering Research and Development* (IJMPERD, 31 items), *RSC Advances* (16 items), *Spectrochimica Acta Part A, Molecular and Biomolecular Spectroscopy* (14 items), *Saudi Journal of Anaesthesia* (12 items), *SCIENTIA Series A: Mathematical Sciences* (12 items), *2010 2nd International Conference on Computer Engineering and Technology* (11 items), *Journal of Hazardous Materials* (11 items), *Journal of Fundamental and Applied Sciences* (10 items), and *The Scientific World Journal* (10 items).

These 816 source titles are published in 40 different countries, with the majority from four countries: the United States (328 source titles), the United Kingdom (251 source titles), India (219 source titles), and the Netherlands (217 source titles), followed by Germany (55 source titles), Egypt (24 source titles), Switzerland (22 source titles), New Zealand (13 source titles), France (8 source titles), and Canada & Singapore (7 source titles each). A total of 204 publishers are involved with these 816 source titles, and top ten includes: Elsevier (319 publications), Springer (213 publications), Wolters Kluwer (85 publications), Wiley (65 publications), Taylor and Francis (57 publications), Wolters Kluwer – Medknow (41 publications), Hindawi (40 publications), IEEE: Institute of Electrical and Electronics Engineers (40 publications), Royal Society of Chemistry (RSC, 40 publications), American Society for Biochemistry and Molecular Biology (ASBMB, 37 publications), and PloS (31 publications).

As indicated in the methodology section, this study goes beyond simple statistical analyses by deploying data carpentry methods to merge related datasets (called cross function in OpenRefine) like the Scopus source list, a public data dump of journal metadata from DOAJ, and the Scimago journal ranking dataset for in-depth analyses of the retracted items under study (1,376 items), like journal quartiles ranges, cite scores and impact factors (IF) ranges, SJR and H-index ranges, and coverage of the source titles in global indexes like DOAJ and Scopus. A cross function call from the primary dataset to the DOAJ journal metadata dataset (a public data dump of journal metadata released on September 30th 2022) shows that 274 retracted items (19.91% of items) were published in 134 DOAJ-listed journals (16.42% of 816 source titles). Most of these 274 retracted items belong to Quartile 1 journals (Q1, 113 items, 41.24%) of DOAJ. Surprisingly, 96 journals of these 134 DOAJ-listed journals (71.64%) have plagiarism-based screening policies; 41 journals have DOAJ Seal (30.59%); 93 journals do not charge APC (69.40%), whereas 41 journals have APC (30.59%). A similar cross function call to the Scopus source list dataset (released in August 2022) reveals that 875 retracted items (63.59% of items) were published in 558 Scopus-listed journals (68.38% of 816 source titles). Most of these 875 retracted items belong to 248 Scopus-listed Quartile 1 journals (Q1, 430 items, 49.14%). A total of 112 journals out of 816 source

Table 10 — Summary of the qualities of journals for retracted items

Journal Quartile N=1,196 items N=706 sources		Journal H-Index N=1,234 items N=726 sources		Journal Impact Factor N=901 items N=528 sources		Journal SJR N=1,202 items N= 713 sources		Journal CiteScore N=1,179 items N=698 sources
Quartile	Items Sources	H-Index range	Items Sources	IF range	Items Sources	SJR range	Items Sources	JCS range	Items Sources
Q1	580	751+	2	7.51 +	105	7.51 +	14	7.51 +	189
	318		1		32		12		100
Q2	333	501-750	8	5.01-7.50	80	5.01-7.50	6	5.01-7.50	300
	228		6		47		4		145
Q3	242	251-500	117	2.51-5.00	398	2.51-5.00	35	2.51-5.00	302
	129		22		223		17		222
Q4	41	1-250	1107	0.10-2.50	318	0.10-2.50	1147	0.10-2.50	388
	31		697		226		680		231

titles (13.72%) are listed in both DOAJ and Scopus.

It has been found that many authors publish retracted items in quality journals[33,36–40]. The integration of the primary dataset with the Scimago, DOAJ, and Scopus datasets on the basis of the coupling of source title and ISSN as a composite unique key affirms some of the interesting facts about the destinations of the retracted items, particularly journals:

- 1,196 items were able to find places in the reputed journals (86.91% of the retracted items were published in journals having quartile scores (Q1 to Q4) in Scimago database); ii) the majority of the retracted items were published in Q1 journals (42.15% of retracted items appeared in Q1, which is 38.97% of source titles);
- 10 retracted items were published in journals having an H-Index higher than 500, and the list includes journals like *NEJM: The New England Journal of Medicine* (H-Index 1079), *Cell* (H-Index 814), *Lancet* (H-Index 807), and *JAMA: Journal of the American Medical Association* (H-Index 709);
- 65.47% of retracted items (901 items) were published in 528 journals (64.70% of source titles) with impact factors ranging from 0.25 to 202.731 (*The Lancet*), and the majority of retracted items (318 items, 35.29% of 901 items) appearing in 223 source journals with impact factors ranging from 2.51 to 5.00 (42.23% of 528 source titles);
- 1,202 items (87.35% of retracted items) were published in 713 journals with an SJR score (87.37% of source titles); and
- The CiteScore values of the destination journals also tell a similar story – 85.68% of retracted items appeared in 698 journals (85.53% of source titles) that have CiteScores from the Scopus database, with the largest chunk of retracted items (602 items, 51.06% of 1,179 items) appearing in the 367 source journals with CiteScores range from 2.51 to 7.50 (52.57% of 698 sources).

The journals include *The Lancet* (IF: 202.73), *JAMA: Journal of the American Medical Association* (IF: 157.3) and so on. Table 10 and Figure 6 together provide a broad view in relation to the quality of source titles that were destined for the retracted items from India.

Impact and openness of retraction

We retrieved citations and altmetric data for 1,267 retracted items with DOI (out of 1,376 items) from ODbL-based citation data sources: Open Citation Corpus (OCC), Scite, Dimensions (for citations) and Altmetric.com (for altmetric attention scores). All data wrangling activities related to citation and altmetric were carried out concurrently in a snapshot on September 30th 2022. The Scite (scite.ai) response in JSON format requires special mention here – it not only provides total citations but also classifies citations into three groups—supporting, contradicting, and mentioning. Results show that Dimension.ai's citation dataset is the most comprehensive one (83%+ responses) and also give data about recent citations. The Scite citation results show that the number of citations used to support a retracted item is higher than the converse, i.e., cited to contradict or oppose views of a retracted item, and most of the citations just mention it as tacit citation (Table 11).

All three deployed citation data sources provided results of similar nature. Not only around 80% of retracted items have received citations, but more than

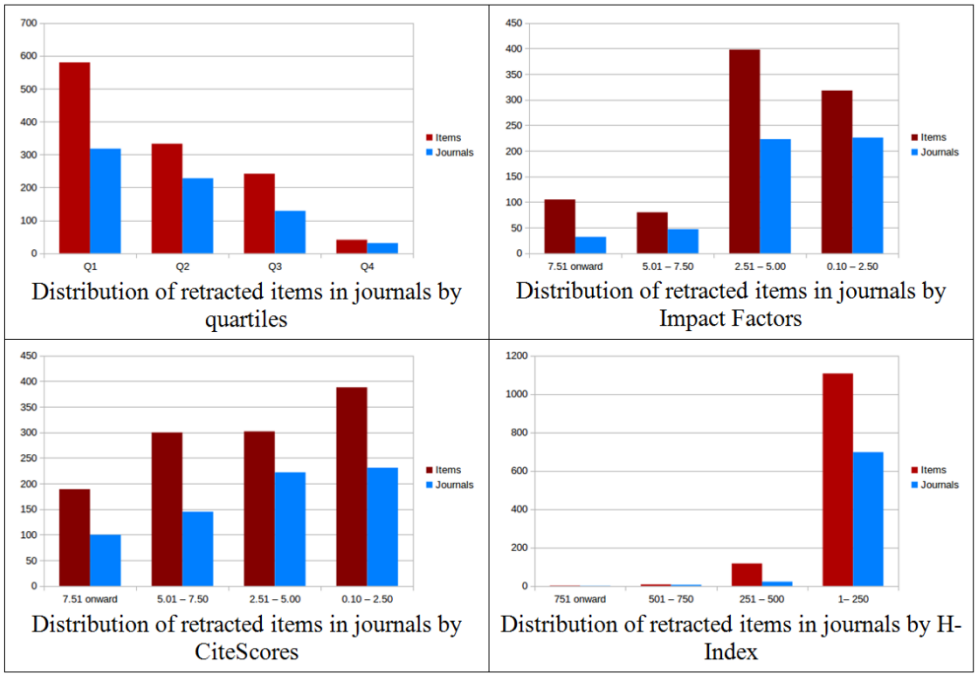

Fig. 6 — Relationships between retraction and journal quality indicators

Table 11 — Citation profiles of retracted items

Citation data sources	Query sent	Responses received	Item with citation >=1	Highest citation by item	Recent citations	Contradicting citations >=1	Supporting citations >=1	Mentioning citations >=1
						(number of retracted items)		
Scite (scite.ai)	1267	1242	910 (71.82%)	1283	NA	49 (3.86%)	264 (20.83%)	895 (70.63%)
Dimensions (dimensions.ai)	1267	1248	1055 (83.26%)	1285	654 (51.61%)	NA	NA	NA
Open Citation Corpus (OCC)	1267	1262	1006 (79.40%)	926	NA	NA	NA	NA

50% of these items were cited in the last two years (in Dimensions.ai, the recent citations value is the number of citations that were received in the last two calendar years) and 33 retracted items have received 100 or more citations. Amazingly, 589 retracted items (46.48% of 1,267 items) have registered their presence in socio-academic web-space and received altmetric attention scores. The highest altmetric attention score achieved by a retracted item related to COVID-19 (DOI: 10.1101/2020.01.30.927871) is 13662.03800000188.

Earlier researchers predicted that delay in retraction is the most visible reason for attracting citations by these fraudulent academic content[22-27], but the dataset related to recent citations as obtained by this study from Dimensions.ai (Table 12) proves that these items receive considerable citations even after their formal retraction, which thereby contradicts the views of the earlier researchers.

The mean citation values for these 1,267 items with DOI from the three deployed citation data sources are: 13.35 (Scite), 16.48 (OCC) and 19.73 (Dimensions). However, this could be an interesting area for future study to explore why retracted items receive citations even after their withdrawal. The REST/API-based content negotiation from Unpaywall reveals the access profiles of 1,242 retracted items out of 1,267 items with DOI. It shows that 584 retracted items are made available in OA routes (46.09% of 1,267 items with DOI), and more than 50% of these open-access retracted items (52.39% to be exact) are published in open-access journals (Gold path). A total of 361 retracted items are available in repositories across the globe (61.81% of 584 OA retracted items) and a total

Table 12 — Recent citations and Altmetric profiles of retracted items

Recent citations range (for 1267 retracted items with DOI, responses received for 1248 items) Source: *Dimensions.ai*	No recent citations	1 to 10	11 to 20	21 to 30	31 to 40	41 to 50	50 +
	(number of retracted items)						
	594	587	38	14	4	1	10 *
	* The highest recent citation is registered as 198 (DOI: 10.4103/0974-1208.82352)						
Altmetric attention score range (for 1267 retracted items with DOI, responses received for 589 items) Source: *altmetric.com*	No altmetric score	0.1 to 10	10.01 to 20	20.01 to 30	30.01 to 40	40.01 to 50	50.01 +
	(number of retracted items)						
	678	477	76	7	6	3	20

of 366 items (62.67% of 584 OA retracted items) are available under an OA license with four major Creative Commons licenses: cc-by (186 items), cc-by-nc-sa (95 items), cc-by-nc-nd (37 items) and cc-by-nc (24 items).

Subject distribution of retraction

We attempted to find patterns in retracted items published from India, by analysing subject categories of these items at two levels: i) in broad disciplines; and ii) on the basis of the major keywords under the broad disciplines. The Retraction Watch database[6] has a set of seven broad categories: (B/T): Business and Technology; (BLS): Basic Life Sciences; (ENV): Environmental Sciences; (HSC): Health Sciences; (HUM): Humanities; (PHY): Physical Sciences; and (SOC): Social Sciences. Each of these broad categories divided into subcategories like (BLS) categories have 19 subcategories under it e.g. (BLS) Agriculture, (BLS) Anatomy/Physiology, (BLS) Biochemistry, (BLS) Biology – Cancer, (BLS) Biology – Cellular, (BLS) Biology – General, (BLS) Biology – Molecular and so on. The primary dataset for this study adopted these broad categories and subcategories for describing the subject content of a retracted item.

This study did not attempt to index the subject content of retracted items on its own; rather, subject keywords for these items were fetched from CrossRef (1,175 responses) and Semantic Scholar (883 responses) via REST/API-based content negotiation based on DOI (1,267 items have DOI). These 1,376 items have been allotted a total of 3,446 subject categories and subcategories by the Retraction Watch database, with the majority of the items having more than one subject category (minimum 1 – 49 items only, maximum 8), and there are 111 distinct values in these 3,446 occurrences of subject categories and subcategories. The highest number of retracted items belongs to the subcategory (BLS) Biochemistry (267 occurrences) and (BLS) Biology – Cellular is a close second (264 occurrences). The other six subcategories that are repeated more than 100 occurrences are: (PHY) Chemistry (179), (BLS) Biology – Molecular (144), (BLS) Microbiology (139), (PHY) Materials Science (135), (BLS) Genetics (117), and (BLS) Plant Biology/Botany (110). These 8 subject categories/subcategories accounted for 785 retracted items (57.04% of 1,376 items). Obviously, a decisive majority of the retracted items belong to different facets of science and technology, but a few retracted items (54 items) have a category stamp from the social sciences like: (SOC) Psychology (12 occurrences), (SOC) Education (11 occurrences), (SOC) Sociology (11 occurrences), (SOC) Communications (7 occurrences), (SOC) Forensics (6 occurrences), and (SOC) Law/Legal Issues (5 occurrences) are all lead labels in the Social Sciences. A density-based bubble chart in Fig. 7 provides a panoramic view of the distribution of subject categories and subcategories in the retracted items.

There are total 3,041 occurrences of subject keywords (238 distinct keywords) from CrossRef to represent content of the 1,175 items. Most of the items are represented by three and four subject keywords (227 instances and 163 instances respectively). The top 10 subject categories for these retracted items are: General Medicine (235 occurrences), Biochemistry (109 occurrences), Molecular Biology (97 occurrences), General Chemistry (88 occurrences), Cell Biology (68 occurrences), Mechanical Engineering (54 occurrences), Organic Chemistry (51 occurrences), Condensed Matter Physics (48 occurrences), General Materials Science (47 occurrences), and Pharmacology (46 occurrences).

As evident, these subject keywords from CrossRef are conforming to the subject categories / subcategories of the Retraction Watch database. Most

Fig. 7 — Subject categories/subcategories (3,446 occurrences) of retracted items (1,376)

of the retracted items belong to science and technology in general but a few items (79 items) are having keywords (34 distinct keywords) from social science category like Renewable Energy & Sustainability (17 occurrences), Management, Monitoring, Policy and Law (9 occurrences), Law (8 occurrences), Health Policy(5 occurrences), Psychiatry and Mental health (5 occurrences), Safety, Risk, Reliability and Quality (5 occurrences), Business and International Management (4 occurrences), Education (4 occurrences), General Social Sciences (4 occurrences), Safety Research (4 occurrences) and Cultural Studies (3 occurrences). Interestingly, two retracted items (DOI: 10.1080/0361526X.2019.1595808 and DOI: 10.1007/s10639-021-10503-5) have keyword library and information science. Figure 8 shows a broad overview of the distribution of subject keywords for the retracted items using a treemap chart based on keyword density.

Institutional and geospatial distribution of retraction

A total of 2,343 instances of retractions involving 964 distinct Indian institutes are associated with 1,376 retracted items under study. These institutes are distributed across the country and range from high schools to universities, remote engineering colleges to IITs, private medical colleges to ICMR research laboratories, corporate R&D units to CSIR institutes. This study has classified these 964 Indian institutes into 10 broad categories on the basis of the nature and importance of these institutes in the country. The category that includes *UGC-affiliated universities, colleges, and IUCs* has registered the highest instances of retraction with 521 items (37.86% of 1,376 items). In fact, top five broad categories of institutes cover 93.31% of retracted items (see Table 13). All India Institute of Medical Sciences, New Delhi; Indian Institute of Technology (ISM), Dhanbad; Bhabha Atomic Research Centre; Indian Institute of Toxicology Research, Lucknow; and Indian Agricultural Research Institute, New Delhi are among the 9 institutes associated with the most retracted items under the respective groups.

The geospatial distribution of all institutional instances of retraction (2,343 instances by 964 Indian institutions) covers 32 states and union territories (out

of 36). The top five states are Tamil Nadu (334 occurrences), Uttar Pradesh (303 occurrences), West Bengal (239 occurrences), Delhi (229 occurrences), and Maharashtra (176 occurrences).

A city-wise analysis shows that the retraction density based ranking for the top five cities is: New Delhi (225 occurrences), Kolkata (137 occurrences),

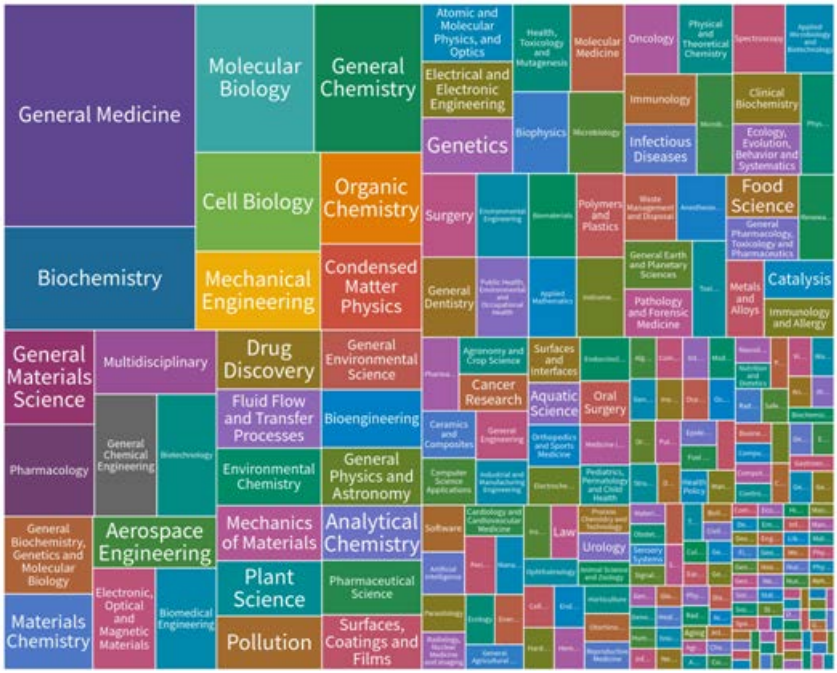

Fig. 8 — Subject keywords (3,041 occurrences) of retracted items (1,175 items)

Table 13 — Institutional categories of retracted items

Category of Institutes [No. of institutes]	Scope of the category	No. of retraction (% of 1,376 items*)	Top institute under the category (with retracted items and %)
UGC-University, Colleges & IUC [250 institutes]	UGC-affiliated public universities, colleges and inter university consortium (IUC)	521 (37.86% of total)	S.V. University, Tirupati (21 items – 4.03% of group total)
Medical Sciences (E & R) [299 institutes]	All medical colleges including AIIMSs and ICMR research institutes (E & R stands for Education and Research)	347 (25.21% of total)	All India Institute of Medical Sciences, New Delhi (28 items – 8.06% of group total)
AICTE-Institutes [187 institutes]	All AICTE-affiliated colleges, universities and other institutes	219 (15.91% of total)	M.Kumarasamy College of Engineering, Karur, Tamil Nadu (31 items – 14.15% of group total)
Elite Institutes of Technology & Research [36 institutes]	Four groups of institutes – IISc, IITs, IIMs and NITs	169 (12.28% of total)	Indian Institute of Technology (ISM) Dhanbad (37 items – 21.89% of group total)
Research & Education Institute (Other categories) [72 institutes]	All other categories of research and educational institutes including Govt. departments	129 (9.37% of total)	Bhabha Atomic Research Centre (14 items - 10.85% of group total)
CSIR-Research Institutes [32 institutes]	All national laboratories affiliated to CSIR	124 (9.01% of total)	Indian Institute of Toxicology Research, Lucknow (22 items – 17.74% of group total)
ICAR-University & Institute [51 institutes]	All institutes affiliated to ICAR including research organizations under ICAR	63 (4.57% of total)	Indian Agricultural Research Institute, New Delhi (6 items – 9.52% of group total)
Private-University & Institute [17 institutes]	All UGC-affiliated private universities, colleges and institutes	55 (3.99% of total)	SASTRA University Thanjavur, Tmail Nadu (11 items – 20% of group total)
Corporate-R&D [33 institutes]	All entities belong to corporate agencies	33 (2.39% of total)	Nutech Mediworld, New Delhi (3 items – 9.09% of group total)

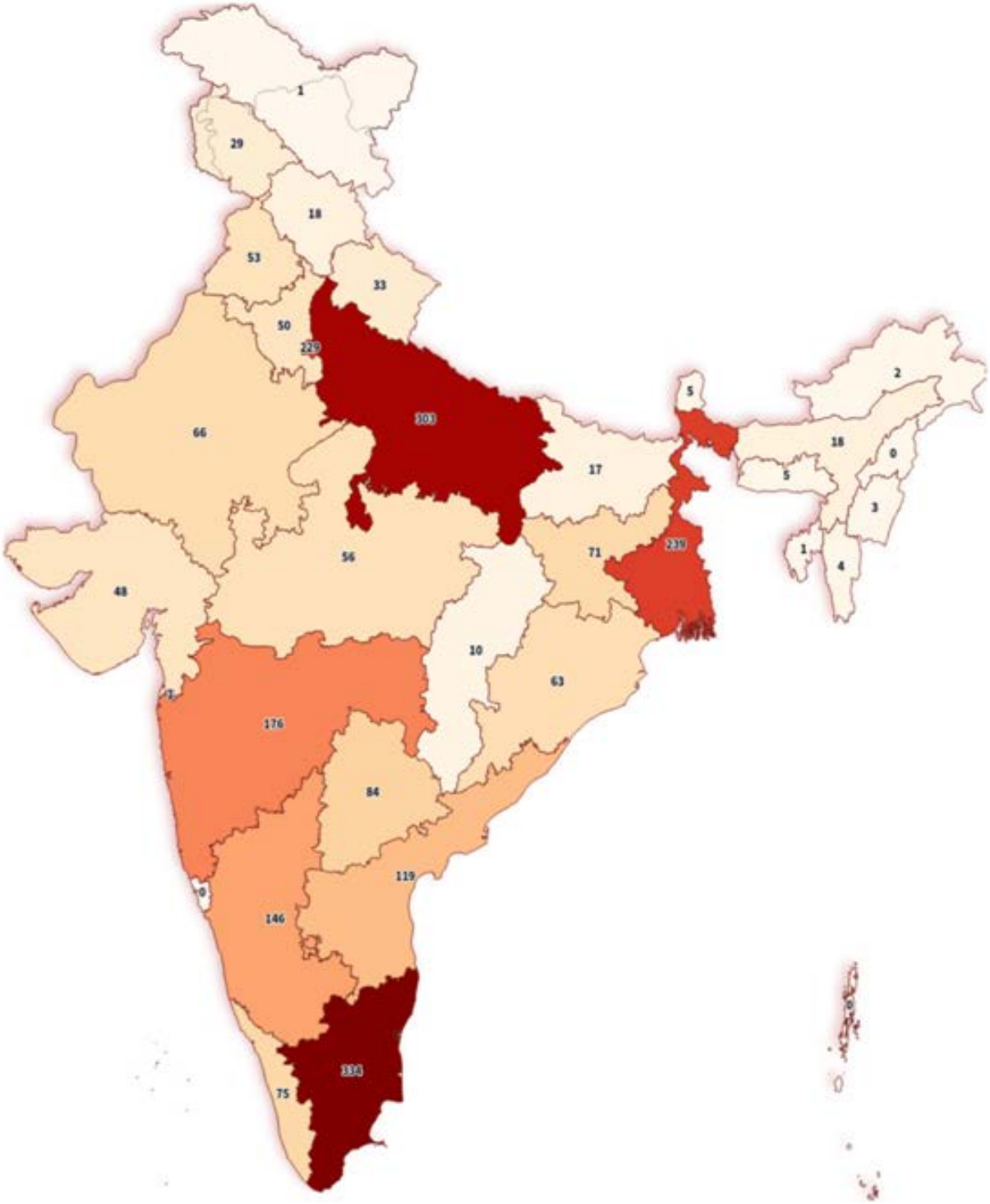

Fig. 9 — Retraction map of India

Individual [4 persons]	Researchers without affiliation	4 (0.29% of total)	–

* One retracted item may be connected to many categories when there is inter-category collaboration. One item, for instance, is linked to eight institutions: two from UGC-University, Colleges & IUC, two from Elite Institutes of Technology & Research, and five from the category of Medical Sciences (E & R).

Lucknow (107 occurrences), Chennai (97 occurrences) and Bengaluru (70 occurrences). This

Fig. 10 — Reasons for retraction in India

research study has attempted to create a retraction map for Indian states and union territories (Fig. 9). This type of map is called a choropleth—a statistical thematic map employs colour that corresponds to an overall summary of a value attribute (here 2,343 instances of retraction) within geographical enumeration units (here 32 states and union territories of India).

Reasons for retractions

The ICAI fundamental values of academic integrity[41] and a Retraction Watch blog[42] that compiles a summary of reasons for retraction, point to a set of causes containing more than 90 possible reasons for retraction, of which plagiarism has only four related reasons: plagiarism of article, plagiarism of data, plagiarism of image, and plagiarism of text. But if we look at the academic integrity guides produced by UGC[4] and AICTE[5], they only talk about plagiarism and do not even bother to define the concept of plagiarism, what are the different categories of plagiarism, and what is the reason for different threshold tolerance values for plagiarism.

However, a review of the factors that led to the retraction of 1,376 items under study provides answers to a few questions that have not yet been adequately addressed. A total of 507 items (36.84%) are associated with only one reason for retraction, but the remaining 869 items are associated with two to eight reasons. Altogether, there are 3,018 instances of reasons involving 84 distinct reasons for retraction of these 1,376 items (one reason may be attached to more than one item, and one item may have more than one reason). This research study has grouped these 84 different reasons into the following ordered array of 12 broad heads for a better understanding of the situation: Text manipulation (741 occurrences), Image manipulation (333 occurrences), Data manipulation (309 occurrences), Errors & Mistakes (285 occurrences), Authorship manipulation (247 occurrences), Notices & Misconduct (242 occurrences), Copyrights, Permission & Conflicts (202 occurrences), Investigations (180 occurrences),

Result manipulation (173 occurrences), Miscellaneous issues (146 occurrences), Withdrawn & Correction (132 occurrences) and Review & Publication issues (28 occurrences).

A summary of the broad groups of reason-sets and the most visible reasons under those groups is illustrated in Fig. 10. It shows that the text manipulation (24.55%) is still the largest reason-set conforming to the result of an earlier research[31], but image manipulation and data manipulation (together these two reason-sets adding to 21.27%) are rapidly catching up. The errors & mistakes group mostly comprising unintentional erroneousness comes next with 9.44% of total occurrences. It is surprising to know from this analysis that retraction occurred for reasons like 'Cites Retracted Work' (5 occurrences) and 'Salami Slicing' (refers to publication of two or more works that grew out of a single study using the same population, methodology, and premise).

Conclusion

Academic integrity policies (along with their subsequent amendments) as developed in India lack necessary supportive datasets, and without data, these are mere opinions, not policy documents. For example, it is difficult for young scholars to understand the scope of plagiarism or the allowed threshold values in the context of the UGC regulation of academic integrity[4]. It has failed to communicate to scholars that plagiarism of text (surprisingly, it is silent about data plagiarism and image plagiarism) is one of the many reasons for retraction (the most serious academic discredit in the professional career of a young scholar), not the only one.

Five items in this dataset, for example, have been retracted due to citations to previously retracted research works (see the classic case for this retraction DOI: *10.1001/jama.2016.18134*). This kind of situation occurs only because of the lack of awareness on the part of the scholars. This research study aims to bridge the gap between the academic integrity policies and the ground realities with the help of this data-intensive study that first developed a comprehensive primary dataset of 1,376 retracted items in the time frame of 75 years (1947–2021), reconfirmed every instance of recorded retraction, and then enhanced the dataset values through integration with an array of related external datasets by using data carpentry methods, tools, and techniques.

References

1 crossref-retractions, (2019), Available at https://github.com/open-retractions/crossref-retractions (Accessed on 25 September 2022).

2 bionode-ncbi, (2022), Available at https://github.com/bionode/bionode-ncbi (Accessed on 24 September 2022).

3 Open retractions, (2022), Available at https://github.com/open-retractions/open-retractions (Accessed on 30 September 2022).

4 The Gazette of India: Extraordinary [PART III—SEC. 4], (31 July 2018), Available at https://www.ugc.ac.in/pdfnews/7771545_academic-integrity-Regulation2018.pdf (Accessed on 20 September 2022).

5 Promotion of academic integrity and excellence and prevention of plagiarism, Available at https://www.aicte-india.org/sites/default/files/Promotion of academic integrity and excellence and Prevention of Plagiarism.PDF (Accessed on 01 September 2022).

6 The Retraction Watch Database. New York: The Center for Scientific Integrity. 2018. ISSN: 2692-465X. Available at http://retractiondatabase.org/ (Accessed on 10 September 2022).

7 Retraction Watch, Publisher retracts 350 papers at once, Retraction Watch Blog, (2022), Available at https://retractionwatch.com/2022/02/23/publisher-retracts-350-papers-at-once/ (Accessed on 11 September 2022).

8 Cabanac G, Labbe C and Magazinov A, Tortured phrases: A dubious writing style emerging in science, Evidence of critical issues affecting established journals. Available at http://arxiv.org/abs/2107.06751 (Accessed on 12 September 2022). DOI: 10.48550/arXiv.2107.06751

9 Vuong Q H, La V P, Ho M T, Vuong T T and Ho M T, Characteristics of retracted articles based on retraction data from online sources through February 2019, *Science Editing*, 7 (1) (2020) 34–44. DOI: 10.6087/kcse.187

10 He T, Retraction of global scientific publications from 2001 to 2010, *Scientometrics*, 96 (2) (2013) 555–561. DOI: 10.1007/s11192-012-0906-3

11 Bar-Ilan J and Halevi G, Temporal characteristics of retracted articles, *Scientometrics*, 116 (3) (2018) 1771–1783. DOI: 10.1007/s11192-018-2802-y

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

13 Li G, Kamel M, Jin Y, Xu MK, Mbuagbaw L, Samaan Z, Levine MA and Thabane L, Exploring the characteristics, global distribution and reasons for retraction of published articles involving human research participants: a literature survey, *Journal of Multidisciplinary Healthcare*, 11 (2018) 39–47. DOI: 10.2147/JMDH.S151745

14 Grieneisen M L and Zhang M, A Comprehensive Survey of Retracted Articles from the Scholarly Literature, von Elm E, editor. *PLoS ONE*, 7 (10) (2012) e44118. DOI: 10.1371/journal.pone.0044118

15 Fanelli D, Costas R, Fang F C, Casadevall A and M. Bik E, Testing Hypotheses on Risk Factors for Scientific Misconduct via Matched-Control Analysis of Papers Containing Problematic Image Duplications, *Science*

and Engineering Ethics, (2019) 771-789. DOI: 10.1007/s11948-018-0023-7

16 Feng L, Yuan J and Yang L, An observation framework for retracted publications in multiple dimensions, *Scientometrics*, 125 (2) (2020) 1445–1457. DOI: 10.1007/s11192-020-03702-3

17 Tang L, Hu G, Sui Y, Yang Y and Cao C, Retraction: the "other face" of research collaboration?, (2020). Available at https://core.ac.uk/display/326244732?source=3 (Accessed on 15 September 2022). DOI: 10.1007/s11948-020-00209-1

18 Mongeon P and Lariviere V, Costly collaborations: The impact of scientific fraud on co-authors' careers, *Journal of the Association for Information Science Technology*, 67 (3) (2016) 535-542. DOI: 10.1002/asi.23421

19 Rathmann J and Rauhut H, Teams prevent misconduct: A study of retracted articles from the Web of Science, (2019). Available at https://core.ac.uk/display/237469952?source=3 (Accessed on 14 September 2022). DOI: 10.5167/uzh-176716

20 Sharma K, Patterns of retractions from 1981-2020 : Does a fraud lead to another fraud?, (2020). Available at https://arxiv.org/abs/2011.13091 (Accessed on 16 September 2022).

21 Fanelli D, Costas R and Lariviere V, Misconduct policies, academic culture and career stage, not gender or pressures to publish, affect scientific integrity, *PLoS ONE*, 10 (6) (2015). DOI: 10.1371/journal.pone.0127556

22 Budd J M, Coble Z and Abritis A, An Investigation of Retracted Articles in the Biomedical Literature, *Proceedings of the Association for Information Science and Technology*, 53 (1) (2016) 1-9. DOI: 10.1002/pra2.2016.14505301055

23 Dinh L, Sarol J, Cheng Y Y, Hsiao T K, Parulian N and Schneider J, Systematic examination of pre- and post-retraction citations, *Proceedings of the Association for Information Science and Technology*, (2019). DOI: 10.1002/pra2.35

24 Teixeira da Silva J A and Bornemann-Cimenti H, Why do some retracted papers continue to be cited?, *Scientometrics*, 110 (1) (2017) 365–370. DOI: 10.1007/s11192-016-2178-9

25 Joob B and Wiwanitkit V, Post retraction citations in context: a comment, *Scientometrics*, 115 (3) (2018) 1291–1292. DOI: 10.1007/s11192-018-2713-y

26 Heibi I and Peroni S, A qualitative and quantitative citation analysis toward retracted articles: a case of study, (2020). Available at https://arxiv.org/abs/2012.11475 (Accessed on 17 September 2022).

27 Frampton G, Woods L and Scott D A, Inconsistent and incomplete retraction of published research: A cross-sectional study on Covid-19 retractions and recommendations to mitigate risks for research, policy and practice, *PLoS ONE*, 16 (10) (2021) e0258935. DOI: 10.1371/journal.pone.0258935

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

29 Bhatt B, A Multi-perspective Analysis of Retractions in Life Sciences, *Scientometrics*, (2020). DOI: 10.1101/2020.04.29.063016

30 Amos K A, The ethics of scholarly publishing: exploring differences in plagiarism and duplicate publication across nations, *Journal of the Medical Library Association*, 102 (2) (2014) 87–91. DOI: 10.3163/1536-5050.102.2.005

31 Elango B, Kozak M and Rajendran P, Analysis of retractions in Indian science, *Scientometrics*, 119 (2) (2019) 1081–1094. DOI: 10.1007/s11192-019-03079-y

32 Wager E, Barbour V, Yentis S and Kleinert S, Retractions: Guidance from the Committee on Publications Ethics (COPE), *Croatian Medical Journal*, 50 (6) (2009) 532-535. DOI: 10.3325/cmj.2009.50.532

33 Steen R G, Retractions in the scientific literature: do authors deliberately commit research fraud?, *Jounal of Medical Ethics*, 37 (2) (2011) 113–117. DOI: 10.1136/jme.2010.038125

34 Santamaria L and Mihaljevic H, Comparison and benchmark of name-to-gender inference services, *PeerJ Computer Science*, 4 (2018) e156. DOI: 10.7717/peerj-cs.156

35 Mukhopadhyay P, Mitra R and Mukhopadhyay M, Library Carpentry: Towards a New Professional Dimension (Part I – Concepts and Case Studies), *SRELS Journal of Information Management*, 58 (2) (2021) 67–80. DOI: 10.17821/srels/2021/v58i2/159969

36 Steen R G, Casadevall A and Fang F C, Why Has the Number of Scientific Retractions Increased?, *PLOS ONE*, 8 (7) (2013) e68397. DOI: 10.1371/journal.pone.0068397

37 van Dalen H P and Henkens K, Intended and unintended consequences of a publish-or-perish culture: A worldwide survey, *Journal of the American Society for Information Science Technology*, 63 (7) (2012) 1282–1293. DOI: 10.1002/asi.22636

38 Van Noorden R, Science publishing: The trouble with retractions, *Nature*, 478 (2011) 26–28. DOI: 10.1038/478026a

39 Cokol M, Lossifov I, Rodriguez-Esteban R and Rzhetsky A, How many scientific papers should be retracted?, *EMBO Reports*, (2007) 422-423. DOI: 10.1038/sj.embor.7400970

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

41 International Center for Academic Integrity, The Fundamental Values of Academic Integrity, ICAI: Fundamental Values of Academic Integrity, (2021), Available at https://academicintegrity.org/images/pdfs/20019_ICAI-Fundamental-Values_R12.pdf (Accessed on 27 September 2022).

42 Retraction Watch, Retraction Watch Database User Guide Appendix B: Reasons, Retraction Watch, (2018), Available at https://retractionwatch.com/retraction-watch-database-user-guide/retraction-watch-database-user-guide-appendix-b-reasons/ (Accessed on 05 September 2022).