Preprint submissions by Indian scientists in arXiv

Vivek Kumar Singh, Satya Swarup Srichandan and Rajesh Piryani

This note analyses the preprint submission patterns of Indian scientists to the popular preprint server arXiv. We analysed research papers published during a five-year period (2014–18) as indexed in the Web of Science and identified how many of these were deposited in arXiv. The discipline-wise distribution of research papers deposited in the repository was also analysed. Results show that overall, only about 3.5% of research papers were deposited in arXiv. The deposits, however, vary across disciplines, ranging from a high of about 23% for physics to a low of 0.4% for agricultural science and biology. We present the overall submission and download statistics for arXiv, and highlight the need for promoting the use of such repositories by the Indian scientific community.

The open-access archive arXiv.org was established in 1991 through collaborative funding as a community-supported resource. It was founded by Paul Ginsparg and is maintained and operated by Cornell University, USA. The server is maintained by a team with guidance from the arXiv Scientific Advisory Board and the arXiv Member Advisory Board. The arXiv repository is funded by Cornell University, the Simons Foundation (https://www.simonsfoundation.org/), member institutions (https://arxiv.org/about/ourmembers) and donors (https://arxiv.org/about/give).

Over a period of time, it has grown to become a 'free distribution service' and an 'open-access archive' for more than 1.5 million scholarly articles (https://arxiv.org). The main subject areas that arXiv covers are: physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science and economics (https://arxiv.org/about). A registered user in arXiv may submit articles that are announced by the server after certain checks. All submissions to arXiv are subject to a moderation process that classifies materials as topical to the subject area and having scholarly value. The submissions can be downloaded free of charge by anyone across the world.

As on 13 July 2020, arXiv has a total of 1,731,310 submissions (https://arxiv.org/stats/monthly_submissions). Figure 1 shows a plot of the number of new submissions per month to arXiv starting from 1991 till 2020. It can be observed that, from about 2000 new submissions per month in 1997, arXiv has grown to receive about 17,000 new submissions per month as in 2020. Ginsparg[1] examined the status of arXiv at the age of 20 years in 2011, and a few years later Van Noorden[2] pointed out that 'arXiv preprint server has hit 1 million articles'. Figure 2 shows the disciplinary distribution of new submissions, indicating that physics, astrophysics, mathematics and computer science constitute majority of these submissions.

Lin *et al.*[3] analysed the arXiv preprints in the area of computer science and showed that there are several reasons why papers are deposited in preprint servers like arXiv. These include: (a) deposits become 'record of priority', (b) they allow 'feedbacks' to authors and (c) deposits create the possibility of submissions becoming 'attention grabbers'. In fact, a careful observation of the statistics of download of papers from arXiv substantiates these claims. According to the download statistics (https://arxiv.org/stats/monthly_downloads) of arXiv, a total of 1,648,280,932 download requests were processed by the server till June 2020. Figure 3 shows the number of downloads per month from arXiv during 1991 to 2020. The arXiv preprint server also includes a membership programme for institutions and libraries, and such members account for more than 75% of institutional downloads. The download statistics also shows the list of top 250 institutions (https://arxiv.org/about/reports/2019_usage) with highest downloads in 2019, which includes two Indian institutions, viz. Physical Research Laboratory, Ahmedabad with 82,991 downloads (0.03% of total downloads) and Indian Institute of Technology Bombay, Mumbai with 73,911 downloads (~0.03% of total downloads).

The arXiv preprint server has thus become a popular service over time with a significant number of papers available on it, which are also being downloaded at a considerably high download rate. However, demographic variations in the usage of arXiv (both submissions and downloads) cannot be ruled out. It is in this context that this note examines how many Indian research papers were deposited in arXiv as preprints before their publication in a journal. Previous studies have found that India has a relatively poor record of open-access availability of its research output[4,5]. Therefore, it would be interesting to analyse whether Indian scientists consider the arXiv preprint server as a preferred mode of providing green open access to their research articles.

Indian research papers in arXiv

We explore the usage of arXiv preprint server by Indian researchers. For this, Indian research output for a period of five years (2014–18) indexed in the Web of Science (WoS) was downloaded to examine how many of these have a preprint available in the arXiv preprint server. It may be noted that a researcher can deposit the preprint in arXiv any time before his/her article is submitted to a journal, or even after the article gets published in a journal. Majority of the journals nowadays clearly provide for such archiving in their copyright transfer statements and in almost all cases researchers are allowed to deposit their article preprints in any institutional or disciplinary repository.

Data and method

The publication records for Indian research output for the period 2014–18 were downloaded from WoS. A total of 377,336 publication records were indexed in WoS for the five-year period, of which 335,503 records had a DOI entry. All these records with DOI were then used as an input for a focused web crawler designed to automatically lookup the arXiv preprint server for the presence of an article. The crawler used DOI and title fields for matching WoS records with submissions in the arXiv preprint server.

Analytical results

First, the matching entries in WoS records and arXiv were identified. Table 1 shows the year-wise data of WoS publication records and the number of records for which a preprint (or post-print) was found in arXiv. It was observed that for all the five years, the percentage of WoS records found in arXiv ranged between 3 and 4, with the five-year average value of 3.52. This is a low number of deposits. However, this percentage of preprint availability was for the entire Indian research output (comprising all the disciplines).

It may be noted that arXiv is more popularly used preprint server by researchers in some specific subject areas. Therefore, it would be more relevant to look at the preprint availability for Indian research output in these selected areas of arXiv. For this, we categorized the publication records into 14 broad disciplinary areas as proposed by Rupika *et al.*[6] These are: agriculture (AGR), art and humanities (AH), biology (BIO), chemistry (CHE), engineering (ENG), environmental science (ENV), geology (GEO), information sciences (INF), materials science (MAR), mathematics (MAT), medical science (MED), multidisciplinary (MUL), physics (PHY) and social science (SS). Such grouping allowed a clear and manageable categorization of publication records classified into 255 subject categories in WoS.

Table 2 presents discipline-wise research output and the number of publication records that have a preprint available in arXiv. It can be seen that for PHY, about 23.36% of the published papers have a preprint available in arXiv. This is followed by MAT with 6.22%, INF with 1.69% and MUL with 1.97% publication records. MAR, ENG and GEO have 3.05%, 0.92% and 0.86% publication records respectively with a preprint available in arXiv. The other seven disciplinary areas, which are not the main areas of arXiv, taken together have a preprint available in arXiv only for 0.42% publications. Figure 4 shows the year-wise trend of publication records for different disciplines that have a preprint available in arXiv. It can be seen that PHY, which was also the initial focus area of arXiv, has between 20% and 26% publication records with a preprint available in arXiv.

Conclusion

We analysed Indian research output for 2014–18 obtained from WoS and the proportion of these research papers having a preprint available in arXiv. It was found that only about 3.52% of total research output has a preprint available in arXiv. These values are higher for disciplines like PHY (23.36%), MAT (6.22%) and MAR (3.05%), which are also the main focus areas of arXiv. However, these proportions indicate the limited use of arXiv preprint repository by the Indian scientific community.

arXiv is a popular and powerful source for disseminating preprints (as well as post-prints) of a large number of research articles, many of which may be behind a paywall. Thus, researchers can make their work available and accessible to the large scientific community by depositing their papers (preprints or post-prints) in arXiv. Also, deposits in arXiv often lead to higher citations and reduced downloads from publisher websites[7].

With regard to Indian research output, it has been found in some previous studies that the major Indian institutional repositories (IRs) get a very low proportion of Indian papers deposited[5,8]. There may be various reasons for this, including the fact that many of the Indian IRs are not indexed by popular web search engines (such as Google). Lack of such discoverability by web search engines make Indian IRs less attractive for Indian researchers to disseminate their work in open access. However, the arXiv preprint server has much higher visibility and is also indexed by popular web search engines. Therefore, Indian researchers can use arXiv as a powerful medium of disseminating their work in an open-access mode. However, the current status of arXiv usage by Indian researchers is not encouraging. More efforts are needed to promote Indian researchers to use arXiv and other similar repositories. These efforts should not only be limited to funding agencies mandating deposits of preprints in IRs or disciplinary repositories, but must also be taken up at the level of institutions, which could make it mandatory for their researchers to submit their papers (preprints or post-prints) to a suitable preprint server or an IR. Such efforts could even go beyond institutions to individuals, which may include encouraging and incentivizing researchers who champion in depositing their work in free and openly accessible repositories (institutional or disciplinary).

1. Ginsparg, P., *Nature*, 2011, **476**(7359), 145–147.
2. Van Noorden, R., *Nature*, 2014; doi:10.1038/nature.2014.16643
3. Lin, J., Yu, Y., Zhou, Y., Zhou, Z. and Shi, X., *Scientometrics*, 2020, **124**, 555–574; https://doi.org/10.1007/s11192-020-03430-8
4. Piryani, R., Dua, J. and Singh, V. K., *Curr. Sci.*, 2019, **117**(9), 1435–1440.
5. Singh, V. K., Piryani, R. and Srichandan, S. S., *Scientometrics*, 2020, **124**(01), 515–531.
6. Rupika, Uddin, A. and Singh, V. K., *Curr. Sci.*, 2016, **110**(10), 1904–1909.
7. Davis, P. M. and Fromerth, M. J., *Scientometrics*, 2007, **71**(2), 203–215.
8. Kumar, V. and Mahesh, G., *Curr. Sci.*, 2017, **112**(2), 210–212.

Vivek Kumar Singh and Satya Swarup Srichandan are in the Department of Computer Science, Banaras Hindu University, Varanasi 221 005, India; Rajesh Piryani is in the Department of Computer Science, South Asian University, New Delhi 110 021, India.
**e-mail:*

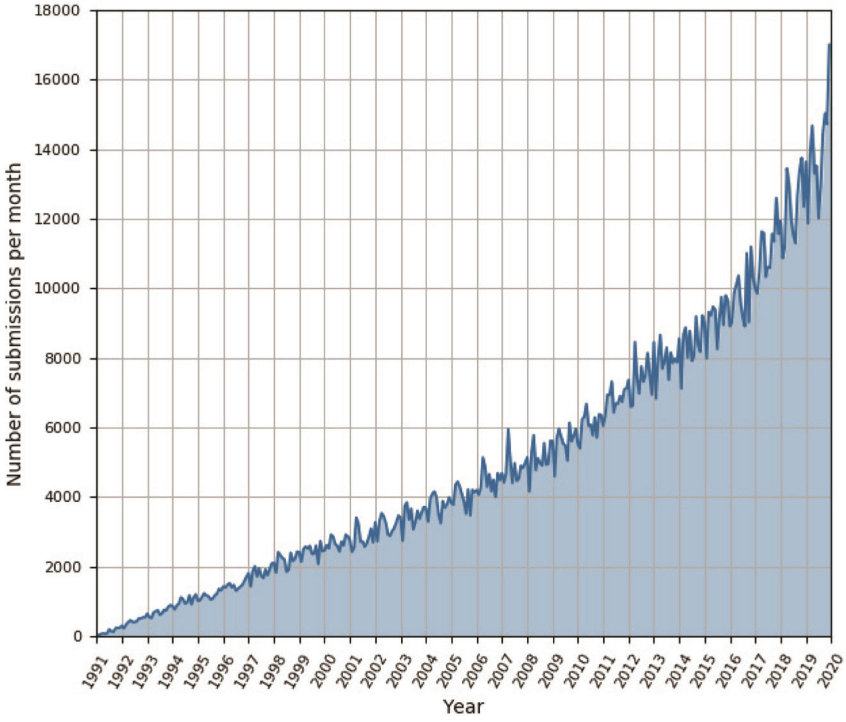

Figure 1. New submissions in arXiv (data from arXiv.org).

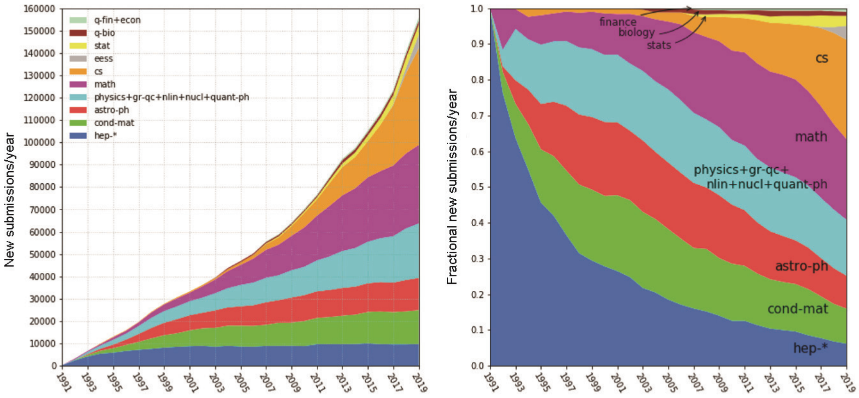

Figure 2. Discipline-wise new submissions in arXiv (courtesy: arXiv.org).

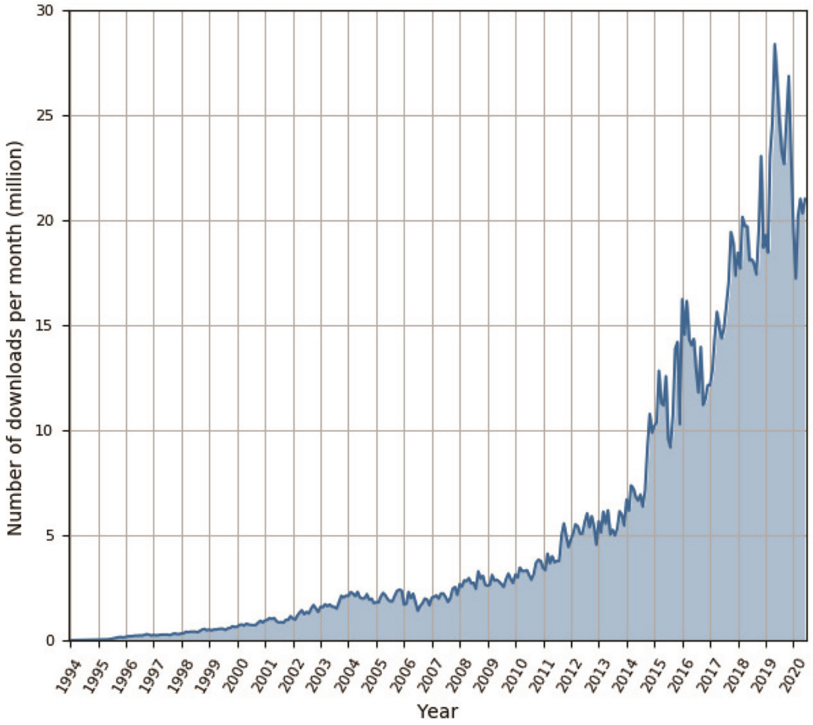

Figure 3. Downloads from arXiv (data from arXiv.org).

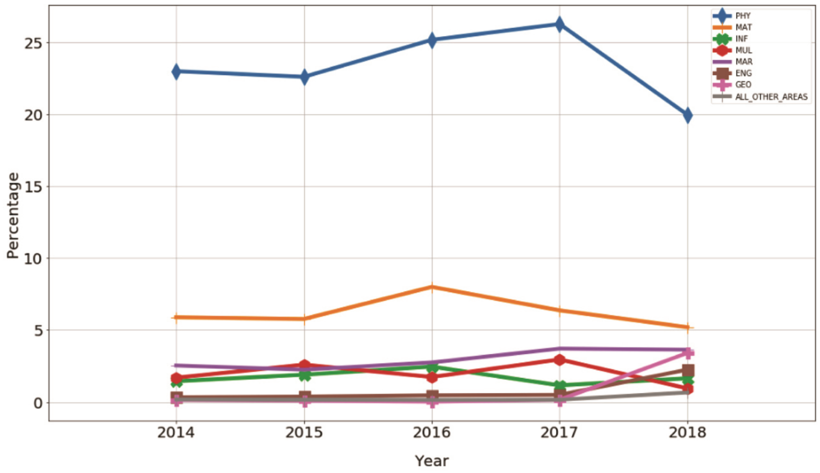

Figure 4. Discipline-wise percentage of Indian articles in arXiv.

Table 1. Year-wise publication records from India and their arXiv presence

Year	Web of Science (WoS) records	WoS records with DOI	Records found on arXiv*
2014	67,575	58,310	2155 (3.70%)
2015	70,685	61,695	1980 (3.21%)
2016	76,530	67,857	2387 (3.52%)
2017	78,532	70,861	2615 (3.69%)
2018	84,014	76,780	2667 (3.47%)
Total	377,336	335,503	11,804 (3.52%)

*Percentage is calculated with respect to the number of articles in WoS having a DOI, i.e. column 3.

Table 2. Discipline-wise count of Indian articles in arXiv (main arXiv disciplines illustrated; publication data are from WoS for period 2014–2018)

Discipline	No. of articles in WoS	No. of articles in arXiv	Percentage
Main disciplines of arXiv			
PHY	38,710	9042	23.36
MAT	7604	473	6.22
INF	15,022	254	1.69
MUL	10,158	200	1.97
Other related disciplines			
MAR	27,756	846	3.05
ENG	31,596	290	0.92
GEO	15,473	133	0.86
All other disciplines			
AGR, AH, BIO, CHE, ENV, MED, SS disciplines (taken together)	195,547	533	0.42