Research Article

Exploring the Publication Metadata Fields in Web of Science, Scopus and Dimensions: Possibilities and Ease of doing Scientometric Analysis

Prashasti Singh[1], Vivek Kumar Singh[2,3], Anurag Kanaujia[3,4,*]

[1]Department of Computer Science, Shri Ram College of Commerce, University of Delhi, Delhi, INDIA.
[2]Department of Computer Science, University of Delhi, Delhi, INDIA.
[3]Department of Computer Science, Banaras Hindu University, Varanasi, Uttar Pradesh, INDIA.
[4]Delhi School of Analytics, University of Delhi, Delhi, INDIA.

ABSTRACT

Publication of a large number of research papers during the last few decades motivated creation of scholarly databases for indexing the publications and recording citations. The publication metadata fields in scholarly databases are now retrieved for various purposes ranging from information search and retrieval to research evaluation. Traditionally, Web of Science and Scopus have been major database sources used. However, with the creation of newer databases like Dimensions, the choice has expanded further. The coverage and citation data of the major databases have been compared in many previous studies. However, there is no existing work on comparison of metadata fields provided by the scholarly databases and the impact that metadata fields provided may have on scientometric research. This work, therefore, attempts to bridge this research gap by comparing the metadata fields contained in the data downloaded through the User Interface (UI) based search from three major scholarly databases- Web of Science, Scopus and Dimensions. The effect of existence or absence of a metadata field in a database on the possibilities and the ease of doing scientometric analysis is explored. The findings are useful for scientometric researchers, practitioners and database managers.

Keywords: Dimensions, Metadata, Scholarly Databases, Scientometrics, Scopus, Web of Science.

INTRODUCTION

The massive growth in scholarly outputs during the last few decades has resulted in creation of several scholarly databases to index the outputs. The Web of Science, Scopus, Google Scholar etc. are few prominent scholarly databases with significant use. Dimensions is a relatively recent addition to the available scholarly database resource pool and has attracted a lot of attention from researchers and practitioners. These databases are used for several purposes ranging from article retrieval to research assessment. Traditionally, research assessment exercises in scientometric studies have more frequently drawn research output data from one of the two well-known scholarly databases-Web of Science or Scopus. The research assessment exercises use metadata stored in the databases in several ways. The scientometric research assessment exercises comprise different metrics such as author-level metrics, publication-citation trends,

national/international collaboration patterns, open access status of research, social media coverage, citation network, journal-level metrics, funding status of institutions/countries etc., In order to compute these metrics and to do scientometric analysis, one needs to obtain metadata for selected research publications from a scholarly database.

The access to metadata of a database usually requires some kind of subscription. For example, Web of Science and Scopus databases can be accessed through a subscription. These subscriptions are often of different types with varying privileges. While some allow only user interface-based access, few advanced grade subscriptions may allow API based access to the database. The databases allow that the required metadata can be downloaded in different file formats such as csv, excel, BibTex, RIS etc., The appropriate file format is chosen according to the type of analysis as well as according to the software package to be used for analysis and visualization. For example, the BibTex or RIS formats can be easily used with visualisation software like VOS viewer,[1] Pajek,[2] Gephi[3] etc., In case, a more advanced programming language-based analysis is to be done, the csv and excel may be more suitable formats. Thus, there exists a choice in terms of data formats and sometimes also in terms of type of access (UI based

or API-based). Some scholarly databases also provide a database dump to selected institutions with regular refreshes. Dimensions database has also provided a Google Big Query based access to its data. However, the user-interface based access is the most Widely available and used route of access for all these databases. In fact, the researchers in the global south may have the user-interface route as the only provided option to access these scholarly databases.

The publication metadata downloaded from the scholarly databases usually contain data classified under various fields. These may include metadata for authors (such as author name, ORCID id, email address, researcher id etc.), organization (such as city/state/country details of the affiliating organization), publication type (such as journal article, conference proceeding, book chapter, letter, erratum, preprint, article in press etc.,) publication source (such as source journal name, book name, conference name, publisher, publisher address, ISSN, eISSN etc.,) citation (such as total citations received, cited references, relative/field citation ratio etc.,) funding details (such as funding organization, funding text, grant ids etc.,) and subject classification (such as broad research area and detailed subject category) etc., Other commonly provided fields include doi, language, publication id, publication year etc.,

Several previous studies have done a comparative analysis of different databases[4-8] and have shown that the databases vary significantly in their coverage. It has also been seen that due to variations in the coverage of databases, the scientometric analysis on data obtained from different databases are found to provide varying evidence.[9] However, there are virtually no studies on comparison of the metadata fields contained in the databases and the impact that such variations may have on the scientometric research exercises. This study, therefore, attempts to explore the variation in metadata fields provided by the three major scholarly databases and to analyse what impact such variation may have on the possibilities and ease of doing scientometric analysis. More precisely, the study attempts to answer the following research questions:

RQ1: What are the major metadata fields in data obtained from UI based search in Web of Science, Scopus and Dimensions and how much do the metadata fields differ across these databases?

RQ2: Do the variation in metadata fields from the three databases result in different possibilities and ease of doing various scientometric analysis?

The metadata fields provided by the three major scholarly databases- Web of Science, Scopus and Dimensions- are analysed. These three scholarly databases are considered as they are the major curated databases being used at present and are considered to be reliable sources of scholarly metadata. Further, they all provide similar ways of subscription-based access, often with different privileges associated with different types of access

provided. For the reasons of uniformity of analysis, we have used the metadata fields provided by the user interface-based access of the three databases for analysis.

The study is not only novel and unique in nature but also important and useful for scientometric researchers and practitioners. The analysis of metadata fields provided by the databases informs what scientometric analysis can or cannot be done with data from a given database. The analysis also informs about the ease of doing specific scientometric analysis if the data is taken from a particular database. By 'ease', we mainly refer to the amount of manual and computational effort required in obtaining and analyzing the metadata for different kinds of analysis. Using metadata from a particular database decides to a large extent, how different kinds of scientometric analysis can be performed. Thus, the researchers and practitioners interested in a specific scientometric analysis can use the knowledge and select a suitable database as data source for the analysis. Further, the findings can be used by database designers and managers in improving the structure, organization and access of the databases.

A BRIEF OVERVIEW OF THE THREE SCHOLARLY DATABASES

This section provides a very brief overview of the three databases (Web of Science, Scopus and Dimensions). For a more detailed information, articles in the special issue of Quantitative Science Studies[1] may be referred.[10-12]

Web of Science

Web of Science, the oldest among the three scholarly databases, has originated from the work on citation index created by Eugene Garfield from the Institute of Scientific Information (ISI) in 1955. Currently, Web of Science is owned by Clarivate Analytics[2]. As per the latest data[3], around 85.9 million scholarly data and 1.9 billion cited references (dating back to 1900) across 254 subject-disciplines is covered by the Web of Science Core Collection. The Science Citation Index Expanded (SCIE) indexes 9,200 journals across 178 scientific disciplines comprising of total 53 million records and 1.18 billion cited references; the Social Sciences Citation Index (SSCI) indexes 3,400 journals across 58 social sciences disciplines comprising of total 9.37 million records and 122 million cited references; and the Arts & Humanities Citation Index (AHCI) indexes 1,800 journals across 28 Arts & Humanities disciplines comprising of total 4.9 million records and 33.4 million cited references.

Web of Science provides a subscription-based access for search and retrieval of data. The data search on the UI interface of Web of Science includes Basic Search, Advanced Search, Author Search and Cited References Search. In the Advanced Search tab, query formulation using Web of Science search tags and Boolean operators is enabled. The data can be downloaded in various formats such as, tab delimited, excel, plain text, RIS,

BibTex formats etc. Moreover, sorting search results via certain options is also available, for example the results of a query can be sorted in ascending/descending order of publication date, highest citations received, relevance etc. There is also a provision of API based access to the database, however it is not provided as part of the most used standard user interface-based access. Further, the advanced access routes are often not available as part of standard subscription in many developing countries.

Scopus

Scopus database, a product of Elsevier, was created in 2004. It covers scientific journals, books, conference proceedings etc. which are selected through a process of content selection followed by continuous re-evaluation. Unlike Web of Science, it has a single citation index, covering journal and conference articles in different subject areas. The Scopus content coverage guide[4] indicates that it contains a total of over 77.8 million core records. As of now, publication records from the year 1788 onwards are covered, with an addition of approximately 3 million records every year. The recently updated content coverage of Scopus shows that it comprises 25,100 active journal titles, more than 210,000 books from 5000 international publishers and over 9.8 million conference papers from over 120,000 worldwide events.

Scopus platform allows data access by Search (Basic and Advanced), Discover and Analyse options. The Basic Search option contains Documents, Authors, Researchers Discovery and Affiliations search while the Advanced Document search enables formulation of a query using Scopus field codes and Boolean operators to search in the Scopus database. The feature for sorting the search results is also available on Scopus platform, for example the results of a query can be sorted in ascending/descending order of publication date, highest citations received, relevance etc. The Discover option enables users to identify collaborators, research organizations with respect to research output, finding related publication data through various metrics such as author keywords, shared references etc., The Analyse option is a tool to track citations, assessment of search results on criteria such as country wise, affiliation wise, research area wise distribution of resultant data. Data can be retrieved from Scopus in various file formats like csv, BibTeX, Plain Text, RIS etc. In addition to access through UI, Scopus also provides API-based access as part of higher-level subscription. In the case of Scopus too, the user-interface based access is the most widely used access route.

Dimensions

Dimensions database was created in 2018 and is a relatively newer database as compared to Web of Science and Scopus. It provides a single platform to access different kinds of research data.[12] Unlike Web of Science and Scopus, Dimensions uses a different approach for sourcing data, with Crossref and PubMed being the "*data spine*". The *bottom up* approach used indicates that the data sourced from Crossref and PubMed is further enhanced by collecting data about affiliations, citations etc. The data enhancements are done through a data enhancement process that takes data from various sources such as DOAJ, initiatives like Open Citations and I4OC, clinical trial registries, openly available public policy data, and other Digital Science companies like Altmetric and IFI Claims. According to the latest data updates[5] Dimensions contains 130 million publications (from around 104,000 journals, 63 preprint servers and 1.5 million books) with about 1.6 billion citation links. Additionally, it indexes data on 6 million grants, 12 million datasets, 226 million online mentions, 883,000 policy mentions, 724,000 clinical trials and 149 million patents.

Dimensions provides different types of access to users such as Dimensions free version, Dimensions Analytics, Dimensions Profiles, Dimensions on Google BigQuery and Dimensions for Life Sciences & Chemistry with different levels of privilege attached with each of them. The user interface-based search enables searching for a particular query in Full Data, Title and Abstract and DOI along with Filters like searching in a particular year or time period, researcher name, research categories, publication type etc. Data from Dimensions can be downloaded in different formats such as csv, excel, BibTex and RIS. Dimensions also allows an API based access which provides more advanced features.

RELATED WORK

Scientometric research assessment exercises have traditionally drawn data from scholarly databases like Web of Science and Scopus. During the last few years many new databases have come up, such as Dimensions, Microsoft Academic (relaunched recently as OpenAlex) etc. The studies on comparison of different databases started with the creation of Scopus and Google Scholar databases.[13-17] While a study[13] compared a German journal list in Social Sciences across Web of Science and Google Scholar databases, another study[14] analysed the overlaps between Web of Science (WoS) and Scopus and some other major scientific databases. Lopez-Illescas *et al.*[15,16] performed an Oncology journal list comparison and their impact on publication-citation trends of countries across Web of Science and Scopus databases while Vieira & Gomes[17] worked on coverage and overlap between Web of Science and Scopus for university domain (for a set of Portuguese universities) for the year 2006. A comprehensive comparison of title coverage, presence of language biases in journal coverage across Web of Science and Scopus was also performed by few studies.[4,18] Web of Science, Scopus & Google Scholar were longitudinally compared on 8 data points on a sample of academicians across different disciplines.[19] Scholarly databases like PubMed, Scopus & Web of Science were analysed on parameters such as coverage, focus and tools using Jordanian authors' data across few disciplines.[20] Similarly, Web of Science

and Scopus were compared on publication type, field of research and language on Norwegian authors' data.[21]

Citation-related studies on scholarly databases has also been an active area of exploration among researchers. A study compared citation patterns comparison across popular scholarly databases like Web of Science, Scopus & Google Scholar on research published by Library and Information Science faculties[22] while another study compared PubMed, Google Scholar, Scopus & Web of Science on citations accrued in the field of Biomedical Research.[23] Citation comparison was performed across Web of Science & Scopus on data in Health Sciences for a Spanish University.[24] Similarly, Web of Science & Google Scholar were compared on citation dataset from 3 UK Business Schools.[25] Macro and micro-level comparison for environmental sciences' journals in South Africa across Web of Science, Scopus & Google Scholar was also performed by certain studies.[26,27] In the same vein, a study[28] compared Web of Science & Google Scholar in development of citation counts in diverse research fields, real growth vs retroactive growth. Other studies[29,30] have performed a comparative analysis of Google Scholar, Web of Science and Scopus for highly cited documents in different subject areas.

The newer databases like Dimensions, Microsoft Academic etc., have also been explored in various respects in the recent studies. Dimensions database was explored as an alternative to the popular scholarly databases like Web of Science and Scopus.[31] Microsoft Academic, Dimensions, Crossref, Web of Science, Scopus & Google Scholar were explored and compared on publication-citation trends in a single academic and six top journals in Business & Economics.[5] A bibliographic comparison of Microsoft Academic, Web of Science & Scopus at institutional level considering data from 15 universities was performed[32] while Dimensions, Crossref, Web of Science & Scopus were compared through publication record match.[33] A comparative analysis of citations to English-language highly cited documents from 252 subject categories in Google Scholar, Microsoft Academic, Dimensions, Open Citations' COCI, Web of Science & Scopus was performed in a study[29] while another study[7] performed a direct as well as pair-wise comparison of Dimensions, Crossref, Microsoft Academic, Web of Science databases with Scopus database on scientific articles. A comprehensive journal coverage comparison using the Master Journal Lists of 3 popular academic databases Web of Science, Scopus and Dimensions was performed in a study by Singh *et al.*[34] A recent work[35] analysed the completeness of author-address links in the Web of Science database during the years 2000 & 2020.

Despite there being several previous studies on various aspects of scholarly databases, the metadata fields provided by these databases have not been analysed. The structure and availability of metadata decides what kind of scientometric analysis can be done and with what ease. Given the fact that the possibilities and ease of doing scientometric assessment is likely to be dependent on the metadata provided by the scholarly database, it is very important that an analysis of the metadata provided by various databases be done. However, to the best of our knowledge there are no existing studies on comparison of metadata fields provided by the three databases and there lies the research gap that this study attempts to bridge.

DATA AND METHOD

The analysis being focused on the metadata fields, the metadata fields in the data downloaded from the three databases, through the user interface route, were identified. For this purpose, the metadata fields were obtained by downloading some publication records by formulating queries in the UI interface of the respective database. The publication metadata fields in Web of Science, Scopus and Dimensions remain the same whether data is downloaded for any search topic, author, journal, particular year or range of years, institution, or a country (this was verified by doing multiple search queries). Therefore, a query on the topic "Artificial Intelligence" was made in all the three databases. In the UI interface of Web of Science database, the search query was TS= "artificial intelligence", where TS stands for Topic Search. Similarly, the search query in Scopus was TITLE-ABS ("artificial intelligence") and in the Dimensions database, the term "artificial intelligence" was searched in the UI interface by limiting search to title and abstract. The publication metadata for the first 200 records was downloaded for each case. From the downloaded records, the metadata fields provided were identified.

The data downloaded from Web of Science contained data classified under 71 metadata fields. The data downloaded from Scopus comprised 46 metadata fields, while the publication data downloaded from Dimensions comprised 55 metadata fields. Some examples of metadata fields in Web of Science are- PT (for Publication Type), DOI (for Digital Object Identifier for a publication), DT (for Document Type), AU (for Authors), C1 (for Address of the Author's affiliating organization), PY (for Publication year of a document) etc. Scopus data comprised metadata fields like Authors (Authors of the publication), Title (Title of the publication), Source title (Journal in which the document is published), DOI (Digital Object Identifier of the publication), Cited by (Total no. of citations received by the publication) etc. The Dimension database has metadata fields like Rank (Relevance ranking of the publication), Publication ID (unique ID provided by Dimensions to each publication), Title (Title of the publication), Abstract (Abstract of the publication), PubYear (Publication year of the document), Open Access (Open Access status of the publication), Author (Authors of the publication) etc.

The metadata fields identified in the three databases were then grouped into different categories for further analysis. The major categories created include metadata for article details, author details, research organization details, publication source details,

citation and usage details, funding acknowledgement details, textual details, open access details, conference details, subject category details. For each of the categories, the metadata fields provided by the three databases were identified. Few records were found common to the three databases, one of those which had maximum populated metadata fields was identified as an example to differentiate amongst the type of data populated in those metadata fields. For this purpose, we used a publication record with DOI '10.1016/j.tourman.2023.104835' that was found indexed with maximum populated metadata fields in the three databases. Thereafter, the possibilities of scientometric analysis that can be done from the available metadata fields in each of the three databases were identified. Subsequently, the ease of doing different scientometric analysis from the available metadata fields of various databases were explored. The Figure 1 provides an illustration of the data and steps of the analysis.

ANALYSIS AND DISCUSSION

The metadata fields provided in the data downloaded from user interface-based search in the three databases Web of Science were identified and categorised into different groups. Around 71 metadata fields were present in the data downloaded from Web of Science, for example, PT (for Publication Type), DOI (for Digital Object Identifier for a publication), DT (for Document Type), AU (for Authors), C1 (for Address of the Author's affiliating organization), PY (for Publication year of a document), AB (for Abstract of a publication), TI (for Title of a publication), LA (for language of the publication), WE (citation index under which the publishing journal is indexed, whether SCIE, SSCI etc.) etc. The complete set of metadata fields provided in the Web of Science UI downloads is shown in Figure 2.

Around 46 metadata fields were present in the data downloaded from Scopus, for example, Authors (Authors of the publication), Title (Title of the publication), Source title (Journal in which the document is published), DOI (Digital Object Identifier of the publication), Cited by (Total no. of citations received by the publication) etc. The complete set of metadata fields found in data downloaded from Scopus UI are shown in Figure 3.

The publication data downloaded from Dimensions through UI search comprises of 55 fields, such as Rank (Relevance ranking of the publication), Publication ID (unique ID provided by Dimensions to each publication), Title (Title of the publication), Abstract (Abstract of the publication), PubYear (Publication year of the document), Open Access (Open Access status of the publication), Author (Authors of the publication) etc. The metadata for obtaining Cited References of the publication can be downloaded by exporting data for bibliometric mapping in csv format (option available in Dimensions data export). The complete set of metadata fields provided in data downloaded from Dimensions user interface-based search is presented in Figure 4.

Metadata for Article Details

The article metadata in databases provides information about article type, DOI, publication date and year etc. Table 1 presents
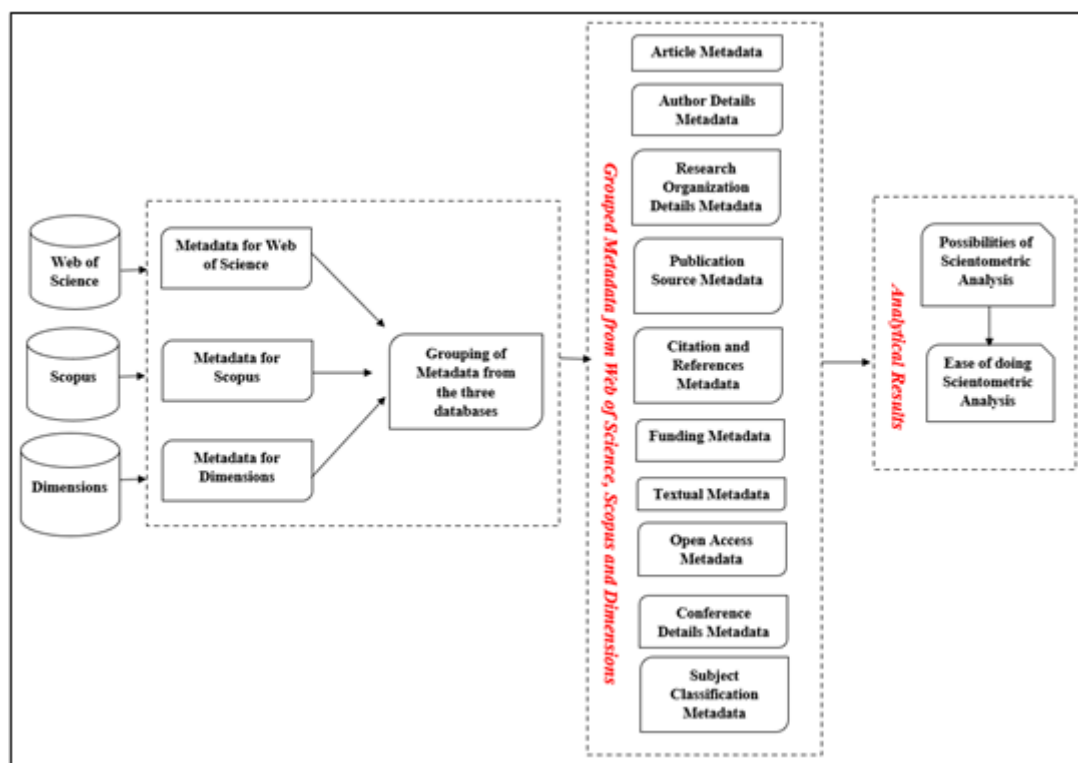

Figure 1: Methodological steps of analysis.

PT, AU, BA, BE, GP, AF, BF, CA, TI, SO, SE, BS, LA, DT, CT, CY, CL, SP, HO, DE, ID, AB, C1, C3, RP, EM, RI, OI, FU, FP, FX, CR, NR, TC, Z9, U1, U2, PU, PI, PA, SN, EI, BN, J9, JI, PD, PY, VL, IS, PN, SU, SI, MA, BP, EP, AR, DI, DL, D2, EA, PG, WC, WE, SC, GA, PM, OA, HC, HP, DA, UT

Figure 2: Metadata Field tags in Web of Science.

Authors, Author full names, Author(s) ID, Title, Year, Source title, Volume, Issue, Art. No., Page start, Page end, Page count, Cited by, DOI, Link, Affiliations, Authors with affiliations, Abstract, Author Keywords, Index Keywords, Molecular Sequence Numbers, Chemicals/CAS, Tradenames, Manufacturers, Funding Details, Funding Texts, References, Correspondence Address, Editors, Publisher, Sponsors, Conference name, Conference date, Conference location, Conference code, ISSN, ISBN, CODEN, PubMed ID, Language of Original Document, Abbreviated Source Title, Document Type, Publication Stage, Open Access, Source, EID

Figure 3: Metadata Fields in Scopus.

Rank, Publication ID, DOI, PMID, PMCID, ISBN, Title, Abstract, Acknowledgements, Funding, Source title, Anthology title, Book editors, Publisher, ISSN, MeSH terms, Publication date, PubYear, Publication date (online), Publication date (print), Volume, Issue, Pagination, Open Access, Publication Type, Document Type, Authors, Authors (Raw Affiliation), Corresponding Authors, Authors Affiliations, Research Organizations - standardized, GRID IDs, City of standardized research organization, State of standardized research organization, Country of standardized research organization, Funder, Funder Group, Funder Country, 'UIDs of supporting grants, Supporting Grants, Times cited, Recent citations, RCR, FCR, Altmetric, Source Linkout, Dimensions URL, Fields of Research (ANZSRC 2020), RCDC Categories, HRCS HC Categories, HRCS RAC Categories, Cancer Types, CSO Categories, Units of Assessment, Sustainable Development Goals

Figure 4: Metadata Fields in Dimensions.

the article metadata fields provided by the three databases. It may be observed that a good number of metadata fields provided by the three databases are common. For example, all the three databases provide fields for document or publication type, publication year, and DOI. Considering a common record among the three databases having DOI '10.1016/j.tourman.2023.104385', the publication type field contains data in the format 'Article' ('DT' in Web of Science, 'Document Type' in Scopus and 'Publication Type' in Dimensions). The other document types indexed in Web of Science are of the form 'Article', 'Editorial Material', 'Review', 'Article; Early Access' etc. in Scopus they are of the form 'Article', 'Review', 'Book chapter', 'Conference paper' etc. while in Dimensions they are of the form 'Reference Work', 'Review Article', 'Research article', 'Research chapter' etc. Another important metadata is the one that contains the full text link of a publication record. The link for full text of a publication record can be found indexed in the 'DL' field in Web of Science, 'Link' field in Scopus and 'Dimensions URL' and 'Source Linkout' field in Dimensions. The common record among the three databases

having DOI '10.1016/j.tourman.2023.104385' contained data as 'http://dx.doi.org/10.1016/j.tourman.2023.104835' in the 'DL' field of Web of Science, and as 'https://www.scopus.com/inward/record.uri?eid=2-s2.0-85170429066&doi=10.1016%2fj.tourman.2023.104835&partnerID=40&md5=1ffb042a2f2f21eb2c183d46f825731d' in the 'Source' field of Scopus. In Dimensions, this particular record didn't contain any entry in the 'Source Linkout' field but contained data as 'https://app.dimensions.ai/details/publication/pub.1163984741' in the 'Dimensions URL' field. This link could then be used to find the full text of the publication record.

However, there are also certain fields that are absent in some databases. For example, Language field is not there in the data obtained from Dimensions database, publication date is not there in data from Scopus and Relevance Rank is not there in the data from Web of Science and Scopus both. The difference in the metadata availability can impact the kind of scientometric analysis possible and/ or the ease of doing such scientometric analysis. Few such cases are discussed next.

Table 1: Metadata Fields for Article Details

Metadata Field	Web of Science	Scopus	Dimensions
Publication Type	PT	-	-
Document Type	DT	Document Type	Publication Type
Language	LA	Language of Original Document	-
Publication Date	PD	-	Publication Date(online), Publication Date(print)
Publication Year	PY	Year	PubYear
Article Number	AR	Art. No.	-
DOI	DI	DOI	DOI
Publication Stage	EA	Publication Stage	-
URL	DL	Link	Source Linkout, Dimensions URL
PubMed ID	PM	PubMed ID	PMID, PMCID
Database ID	UT	EID	Publication ID
Relevance Rank	-	-	Rank

Remarks: No Language details in Dimensions.

The language metadata field of publication records is an important information that can be used to understand the language-wise composition of articles on a research topic or institution or a country. Since this metadata field is not present in data downloaded from user interface-based search of Dimensions database, a language composition analysis cannot be done with Dimensions data. The publication date is another important field, which unfortunately is missing in Scopus. Therefore, any analysis that requires the publication date of articles cannot be done with data downloaded from Scopus database. One such case may be measuring the speed of accumulation of altmetric mentions, which requires information about article publication date. Another interesting point to observe is that the percentage of retracted papers can be obtained from Web of Science and Scopus but not from Dimensions because it doesn't contain Retracted paper count.

Metadata for Author Details

The metadata for author details provided by the three databases is given in Table 2. The 'AU' field in Web of Science, 'Authors' field in both Scopus and Dimensions provides the author names separated by a delimiter. For example, the common record among the three databases having DOI '10.1016/j.tourman.2023.104385' contains data as 'Zhang, JB; Chen, Q; Lu, JD; Wang, XL; Liu, LN; Feng, YQ' ('AU' field in Web of Science); 'Zhang J.; Chen Q.; Lu J.; Wang X.; Liu L.; Feng Y.' ('Authors' field in Scopus) and 'Zhang, Junbo; Chen, Qi; Lu, Jiandong; Wang, Xiaolei; Liu, Luning; Feng, Yuqiang' ('Authors' field in Dimensions). This field in all the three databases can be processed to obtain the count of authors in a publication and hence the authorship patterns for the whole set of publication records on a topic or for an institution. The 'AF' field in

Web of Science and 'Author full names' field in Scopus contain the full name of the authors. For example, the common record among the three databases having DOI '10.1016/j.tourman.2023.104385' contains data as 'Zhang, Junbo; Chen, Qi; Lu, Jiandong; Wang, Xiaolei; Liu, Luning; Feng, Yuqiang' in 'AF' field of Web of Science and 'Zhang, Junbo (58569783100); Chen, Qi (56608316100); Lu, Jiandong (57226095245); Wang, Xiaolei (57188987175); Liu, Luning (56092932300); Feng, Yuqiang (7404544727)' in 'Author full names' field of Scopus. The 'Author(s) ID' field in Scopus contain the Scopus IDs of the authors explicitly which are mentioned in brackets along with author full names in 'Author full names' field of Scopus.

The gender determination scientometric research exercise involves requirement of author first name and affiliation country. In the case of Web of Science, since the 'AU' field contains only initials of authors' first names while 'AF' field contains author full names and not country therefore the 'C1' field in Web of Science which contains the affiliation data has to be used to know the country. These two values together can then be used to determine the gender of the author using a suitable API service. In the case of Scopus, the 'Authors' field as well as 'Authors with Affiliation' field contain the initials of the first name of the authors so these fields cannot be used for gender determination. Instead, 'Author full names' field has to be used to extract the first name of the authors which will require an extra processing operation as the Scopus IDs of the authors are also nested in this field as shown in the example above. Alongside, the 'Affiliations' or the 'Authors with affiliations' field metadata has to be processed to know the country of the affiliating author. Dimensions database provides the full first as well as the last name of the authors in the Authors field. It provides the details of the affiliating organization in the

Table 2: Metadata Fields for Author Details.

Metadata Field	Web of Science	Scopus	Dimensions
Author Name	AU, AF	Authors, Author full names, Author (s) ID	Authors
Email Address	EM	-	-
Researcher ID	RI	-	-
Orcid ID	OI	-	-

Remarks: No ORCID ID, Researcher ID and Email in Scopus and Dimensions.

Authors (Raw Affiliation) field. Moreover, Dimensions database explicitly provides the country of the authors in a dedicated field, Research Org Country field. Thus, Dimensions database provides the author related information in a more direct way which can be easily used for tasks like gender determination. An example on the type of data indexed in research organization metadata is illustrated in the next subsection.

Other major difference in Web of Science and the other two databases is that, while Web of Science provides email address, researcher id and ORCID id of the author explicitly in specific metadata fields, these fields are not explicitly provided by Scopus and Dimensions database in the downloaded data. Thus, any analysis that requires use or linking of ORCID id of authors can only be done with Web of Science data. Also, in order to obtain email ID of the author, the Correspondence Address field of the Scopus data needs to be processed. To illustrate this, considering the common record among the three databases with DOI '10.1016/j.tourman.2023.104385' the 'EM' field contains data as 'liuluning@hit.edu.in' while the 'OI' field contains data as "Liu, Luning/0000-0002-5539-5623' in Web of Science. For the same record, the email address needs to be extracted from the 'Corresponding Address' field in Scopus which contains data as 'L. Liu; School of Management, Harbin Institute of Technology, Harbin, Heilongjiang, 150001, China; email: liuluning@hit.edu.cn'.

Metadata for Research Organization Details

The metadata fields for obtaining the research organization details of a publication record are provided in Table 3. The details consist of organization name, address and some other identifiers assigned by the database. The Web of Science database provides metadata fields 'C1' that contains organization names and addresses (including country) of the authors associated with a publication and 'C3' that explicitly contains the name of the affiliating institutions separated by delimiters. Scopus database provides this information in two fields namely 'Authors with affiliations' and 'Affiliations' that contain author names with affiliations and the affiliation address of the authors respectively. Dimensions database organizes the research organization details in a more structured way. It contains additional fields of standardized organization names, GRID IDs, city, state and country of research organization as separate fields.

In order to illustrate the difference in data populated in the research organization fields among the three databases, we use the common record among the three databases with DOI '10.1016/j.tourman.2023.104385'. For this record, 'C1' field in Web of Science contains data as '[Zhang, Junbo; Lu, Jiandong; Liu, Luning; Feng, Yuqiang] Harbin Inst Technol, Sch Management, Harbin 150001, Heilongjiang, Peoples R China; [Chen, Qi] Dalian Univ Technol, Sch Econ & Management, Dalian 116081, Liaoning, Peoples R China; [Wang, Xiaolei] Univ Int Business & Econ, Sch Informat Technol & Management, Beijing 100029, Peoples R China' and 'C3' field contains data as 'Harbin Institute of Technology; Dalian University of Technology; University of International Business & Economics'. For the same record, Scopus metadata 'Affiliations' contains data as 'School of Management, Harbin Institute of Technology, Heilongjiang, Harbin, 150001, China; School of Economics and Management, Dalian University of Technology, Liaoning, Dalian, 116081, China; School of Information Technology and Management, University of International Business and Economics, Beijing, 100029, China'. In Dimensions database, this record contains data as 'Zhang, Junbo (Harbin Institute of Technology); Chen, Qi (Dalian University of Technology); Lu, Jiandong (Harbin Institute of Technology); Wang, Xiaolei (University of International Business and Economics); Liu, Luning (Harbin Institute of Technology); Feng, Yuqiang (Harbin Institute of Technology)' in 'Authors Affiliations' field and as 'Dalian University of Technology; Harbin Institute of Technology; University of International Business and Economics' in 'Research Organizations – standardized' field. Moreover, from the 'City of standardized research organization' metadata the data was found indexed as 'Dalian; Harbin; Beijing' while 'Country of standardized research organization' metadata contained data as 'China; China; China'. The grid ids of the research organizations for this publication record were found as 'grid.30055.33; grid.19373.3f; grid.443284.d' in the 'GRID IDs' field in Dimensions.

The different metadata fields related to organizations provided by the three databases provide varying possibilities and ease of doing scientometric analysis. The research organization details can be computed with almost equal ease from the three databases. However, the location details such as city, state and country of the research organization are better organized in dedicated metadata fields in Dimensions. Thus, Dimensions

Table 3: Metadata Fields for Research Organization Details.

Metadata Field	Web of Science	Scopus	Dimensions
Organization Address	C1, C3	Affiliation, Authors with Affiliation	Authors (Raw Affiliation), Authors Affiliations
Correspondence Address	-	Correspondence Address	Corresponding Author
Research Organization	-	-	Research Organizations- standardized
Research Org IDs	-	-	GRID IDs
Research Org City	-	-	City of standardized research organization
Research Org State	-	-	State of standardized research organization
Research Org Country	-	-	Country of standardized research organization

provide a direct field for country names associated with each publication record. This allows a direct computation of the number of internationally collaborated papers in any set of publication records. Further, the proportions of bilateral and multilateral collaborated papers could also be computed directly. Taking other data together (such as for subject area or citations) can further allow a direct computation of impact of international collaboration on citation impact of publications in different subject areas. Dimensions database also provides city and state of publication records. This allows a direct computation of city-wise or state-wise research output, including collaboration patterns in them. Similarly, the standardized organization names and GRID IDs provided by Dimensions database allows a better and easier way to handle duplicate records and organization association problems encountered in Web of Science and Scopus databases. The standardized organization names also allow computation of collaboration patterns in different types of institutions, such as university-industry, university-government, university-facility etc. Thus, in terms of the metadata fields provided for the organization details, Dimensions database may be considered superior in organizing and structuring the information. Several types of analysis can be conducted using the Dimensions data in a more direct and easier manner, which is not possible (or difficult) with Web of Science or Scopus provided metadata.

Metadata for Publication Source Details

The publication source details are provided in the metadata fields of all the three databases. However, they are differently structured. All the three databases provide source title, publisher, volume and page number details of the publication records. Web of Science provides additional information of publisher city and publisher address. The ISSN, abbreviated source name and page count information is provided by both Web of Science and Scopus databases but not by the Dimensions database. Further, Dimensions database does not provide the page count information. The metadata for publication source details found

in Web of Science, Scopus and Dimensions is provided in Table 4. Based on the metadata fields provided by the three databases, one may analyse the possibilities and ease of doing scientometric analysis when data is taken from different databases. One such case is that of ISSN, which is not provided by the Dimensions database and hence any analysis involving ISSN cannot be done with Dimensions data. Few studies have made use of page count information to provide an overview of author and article characteristics across OECD subject classification schemes using WoS data[36] while others[37] have studied the relationship between readership and citation patterns based on bibliographic characteristics of documents based on WoS data. The page count information is provided directly in Web of Science and Scopus but not by Dimensions. Therefore, the average page length of articles in a data set can be more directly computed in Web of Science and Scopus as compared to Dimensions, which will need an extra subtraction operation.

Metadata for Citation and Usage

The citation and article usage details of different kinds are provided in the relevant metadata fields of the three databases. All the three databases provide 'times cited' information. Web of Science and Scopus provide 'cited references' details of a publication record, which is not provided by Dimensions database user interface-based search. Though in case of Dimensions, a second level search for citation metadata for selected DOIs can be done by using the option 'Export for bibliometric mapping'. Web of Science exclusively provides usage count information in U1 and U2 fields. It also exclusively provides a cited reference count field. The Dimensions database provides information about recent citations (number of citations accrued by a publication in the last 2 years, reset at the beginning of each calendar year), relative and field weighted citation ratio and the altmetric attention score of a publication record. The Table 5 summarizes the citation and usage related metadata fields provided by the three databases.

The variation in the citation and usage related metadata fields provided by the three databases allow different possibilities and ease of doing scientometric analysis. The first major observation is that no citation network can be created with data provided by Dimensions database as it does not include cited reference details in publication metadata obtained through the UI access route. A separate download for citation information is required in Dimensions followed by DOI matching process with publications. Further, the usage of count-based analysis is possible only with Web of Science database and not with Scopus or Dimensions. U1 reflects the number of times the full text of a publication record has been downloaded in the last 180 days, U2 counts the number of times the full text of a record has been accessed or saved since February 1, 2013. Studies have worked on usage count of research articles.[38,39] Usage count indicator can serve as a metric in determination of latest research fronts.[40] The correlation between usage and citations for a dataset can also be computed with Web of Science data. In addition, Web of Science provides a cited reference count field that can be used to directly compute the average number of references in any dataset being explored, which is otherwise difficult in Scopus and not possible in Dimensions.

The Dimensions database, however, provides an altmetric attention score which can be used to do altmetric analysis of a dataset. This is, however, not possible with Web of Science and Scopus. Altmetric attention score can also be used to compute correlations between citations and altmetrics, as seen in some previous studies.[41-45] Moreover, altmetric attention score data has also been used earlier to observe power law behaviours in altmetric data.[46]

Similarly, Dimensions provides recent citation, RCR (Relative Citation Ratio) and FCR (Field Citation Ratio) fields, all of which can be used for interesting scientometric analysis. There have been widely acknowledged limitations of using raw citation counts, h-index and Impact Factor in determining research impact.[47,48] It is because different fields cite at varying rates and citation counts tend to fluctuate in months to years after a publication has occurred. Thus, metrics such as RCR (Relative Citation Ratio) and FCR (Field Citation Ratio) come into play. RCR is obtained by dividing the number of citations an article

Table 4: Metadata Fields for Publication Source Details.

Metadata Field	Web of Science	Scopus	Dimensions
Publication Source	SO	Source Title	Source title
Publisher Name	PU	Publisher	Publisher
Publisher City	PI	-	-
Publisher Address	PA	-	-
ISSN	SN	ISSN	-
eISSN	EI	-	-
Source Abbreviation	J9, JI	Abbreviated Source Title	-
Volume	VL	Volume	Volume
Issue	IS	Issue	Issue
Beginning Page	BP	Page Start	Pagination
Ending Page	EP	Page End	Pagination
Page Count	PG	Page Count	-

Remarks: No Page Count in Dimensions.

Table 5: Metadata Fields for Citation, Uses and References.

Metadata Field	Web of Science	Scopus	Dimensions
Cited References	CR	References	-
Cited Reference Count	NR	-	-
Times Cited	TC, Z9	Cited by	Times cited
Usage Count	U1, U2	-	-
Recent Citations	-	-	Recent citations
RCR	-	-	RCR
FCR	-	-	FCR
Altmetric	-	-	Altmetric

Remarks: No cited references in Dimensions; No usage count in Scopus and Dimensions.

Table 6: Metadata Fields for Funding Acknowledgements.

Metadata Field	Web of Science	Scopus	Dimensions
Funding Agency and Grant Number	FU, FP	Funding Details	Funder, Funder Group
Funding Text	FX	Funding Text	Acknowledgements
Funding Country		-	Funding Country
Supporting GRANT ID	-	-	Supporting Grants, UIDs of supporting grants

has received by the expected no. of citations of the articles, while FCR is obtained by the average journal citation rate of the articles of the same field. A discussion of citation metrics such as RCR (Relative citation ratio), FWCI (Field weighted citation impact) has been provided in a study.[49] A thorough discussion of such metrics could also be found in a few studies.[50-53] The Dimensions database is thus well-suited for this kind of analysis as it explicitly contains the metadata for both the RCR and FCR values of a publication record.

Metadata for Funding Acknowledgements

The funding and grant details for different publication records is captured by the databases in different metadata fields. The Table 6 presents the metadata fields for funding acknowledgement in the three databases. The Web of Science and Scopus databases both contain funding details and a funding acknowledgement text. The Dimension database contains some additional information in the form of funding country and grant IDs.

Considering the common record among the three databases with DOI '10.1016/j.tourman.2023.104385', in Web of Science the funding details i.e. the Funder Agency name and Grant No. together is provided in 'FU' field like 'National Natural Science Foundation of China [72202037, 72101045, 72034001, 71974044]; Fundamental Research Funds for the Central Universities in UIBE [21QN01]; Fundamental Research Funds in DUT [DUT22RW102]; Heilongjiang Provincial Natural Science Foundation of China [YQ2020G004]; Fundamental Research Funds for the Central Universities [HIT.OCEF.2022054, HIT.HSS. DZ201905]'. The 'FX' field has textual details on acknowledging the funders of the publication as 'This research was supported by the grants from the National Natural Science Foundation of China (72202037, 72101045, 72034001, 71974044), the Fundamental Research Funds for the Central Universities in UIBE (21QN01), the Fundamental Research Funds in DUT (DUT22RW102), Heilongjiang Provincial Natural Science Foundation of China (YQ2020G004) and the Fundamental Research Funds for the Central Universities (HIT.OCEF.2022054 and HIT.HSS. DZ201905)', while the 'FP' field contains only name of the funding agencies as 'National Natural Science Foundation of China (National Natural Science Foundation of China (NSFC)); Fundamental Research Funds for the Central Universities in UIBE; Fundamental Research Funds in DUT; Heilongjiang Provincial Natural Science Foundation of China (Natural Science

Foundation of Heilongjiang Province); Fundamental Research Funds for the Central Universities (Fundamental Research Funds for the Central Universities)'. The funding agency and grant no. details in Scopus for this record were found in the 'Funding Details' field that contained data as 'National Natural Science Foundation of China, NSFC, (71974044, 72034001, 72101045, 72202037); Natural Science Foundation of Heilongjiang Province, (YQ2020G004); Fundamental Research Funds for the Central Universities, (21QN01, DUT22RW102, HIT.HSS.DZ201905, HIT.OCEF.2022054)' while the 'Funding Texts' metadata for the same record contained data as 'This research was supported by the grants from the National Natural Science Foundation of China (72202037, 72101045, 72034001, 71974044), the Fundamental Research Funds for the Central Universities in UIBE (21QN01), the Fundamental Research Funds in DUT (DUT22RW102), Heilongjiang Provincial Natural Science Foundation of China (YQ2020G004) and the Fundamental Research Funds for the Central Universities (HIT.OCEF.2022054 and HIT.HSS. DZ201905).

The grant numbers and the funding agency have to be extracted from these metadata fields in the case of both Web of Science and Scopus databases. The Dimensions database on the other hand contains multiple metadata fields that contain information explicitly like 'Supporting Grants' and 'UIDs of supporting grants' fields; the name of the country that funds in the field 'Funding Country'. For the example record, 'Funder' metadata contained details as "National Natural Science Foundation of China; Ministry of Education of the People's Republic of China" while 'Funding Country' metadata contained data as 'China; China; China'. Similarly, the field 'Supporting Grants' and 'UIDs of supporting grants' field contained data as '71974044' and 'grant.8870988'.

It is known that research funding and incentives to institutions/ universities bolsters publication performance.[54-56] Further, it is expected that publicly funded research output should be openly accessible. For these and other purposes the funding metadata is analysed in several studies. Patterns of funded publication in a given dataset are often analysed along with other relevant fields like citations, subject area, country. While Web of Science and Scopus provide similar metadata and hence similar analysis possible, Dimensions provides a more structured metadata for funding details. Funder Group, funding country and Supporting

Grants are additional fields in Dimensions, which can be used for country details of funding agencies and resultant publications in different subject areas.

Metadata for Textual Details

There are some text related metadata of articles provided by the three databases. These include author keywords, index keywords, abstract and title of the article. The Web of Science and Scopus databases provide all these four fields. However, Dimensions only provides title and abstract fields. The Table 7 summarizes the text-related metadata provided by the three databases. The textual information provided is not only used in retrieval related queries but also for the purposes of creating keyword co-occurrence networks. The keywords (both author and index) have been used in many previous studies for tasks as versatile as keyword density mapping to creation of expertise indices[57] and their applications.[58] Keywords and index keywords are also used to understand the thematic structure of research publications and draw visualization maps. For example, the common record among the three databases with DOI '10.1016/j.tourman.2023.104385', contained data as 'Chatbot; Human-computer interaction; Expectancy violations theory; Emotional expressions; Customer service; Customer satisfaction' in the 'DE' field of Web of Science while it contained data as 'artificial intelligence; boundary condition; computer; consumption behaviour; psychology; social behavior' in the 'Index Keywords' metadata of Scopus. All this thematic analysis is possible with data from Web of Science and Scopus databases. The Dimensions database does not provide keywords in user interface-based search and hence any analysis involving author and index keywords is not possible if the data is obtained from Dimensions database UI access route.

Metadata for Open Access Details

The scholarly databases nowadays also contain information about open access availability of articles. The Table 8 shows the metadata fields provided for the purpose by the three databases. It can be observed that all the three databases provide details about open access availability of a publication. For example, the common record among the three databases with DOI '10.1016/j.tourman.2023.104385', the Web of Science 'OA' field contained data as 'Bronze' while the 'Open Access' field in Scopus and Dimensions contained data as 'All Open Access; Bronze Open Access' and 'Closed', respectively. Thus, the three databases

contain open access status of publication records separated by identifiers like ',' and ';' which can be easily processed to obtain the open access status of publication records.

It is known that institutions across countries are now increasingly participating in open access publishing[59] and that open access availability enhances greater participation in science for various audiences such as authors, researchers, funders etc.,[60] It is particularly beneficial for developing and under-developed countries.[61] There are different routes to open access, such as Gold, Green, Bronze etc.,[62] Owing to the importance associated with open access, several previous studies have tried to measure the extent and the type of open access publishing at the level of a country by using the open access related metadata field.[34,63] Though different studies may have used different databases, there is no difference in the databases regarding the open access related metadata and hence any of them can be used for open access related analysis with the same ease.

Metadata for Conference Details

The scholarly databases include metadata fields for conferences and articles published in their proceedings. The Table 9 summarizes the conference related metadata fields provided by the three databases. The Web of Science provides information about conference title, year, location etc. The Scopus database also provides information related to conference name, date and location. Both databases also provide for a conference code. However, while Web of Science provides the conference year, Scopus provides conference dates. The Dimensions database on the other hand does not provide any conference related metadata. Though it provides a filter for conference search by searching for proceedings, but conference details metadata is not provided. Therefore, any analysis involving conference publications can be done only by using data from Web of Science or Scopus databases and data downloaded from Dimensions UI route cannot be used for this purpose.

Metadata for Subject Categories

The scholarly databases provide for different kinds of subject classification of publication records. The Table 10 summarizes the metadata fields related to subject classification in the three databases. The Web of Science provides a 'WC' field which contains an entry for the Web of Science subject category. There are about 258 subject categories provided by Web of Science.

Table 7: Metadata Fields for Text and Keywords.

Metadata Field	Web of Science	Scopus	Dimensions
Author Keywords	DE	Author Keywords	-
Index Keywords	ID	Index Keywords	-
Abstract	AB	Abstract	Abstract
Title	TI	Title	Title

Remarks: No keywords in Dimensions.

The Web of Science also provides a metadata field 'SC' (Research Area) which provides a top-level subject classification.

For the given example record with DOI '10.1016/j. tourman.2023.104385', Web of Science contained data as 'Environmental Studies; Hospitality, Leisure, Sports & Tourism; Management' in 'WC' category of Web of Science while the 'SC' category in Web of Science contained data as 'Environmental Sciences & Ecology; Social Sciences – Other Topics; Business & Economics' for this record. The Scopus database, however, doesn't provide any metadata on subject classification in the downloaded publication metadata. Although Scopus provides for 27 subject categories, it only allows searching by them in user interface and does not include information about the subject category in the metadata that is downloaded. The Dimensions database provides a metadata field 'Fields of Research (ANZSRC 2020)' which is one of the major subject categories provided in the Dimensions database. The entries for this field comprise of a code and subject name, such as 40 Engineering, 31 Biological Science, 46 Information and Computing Sciences etc. The Fields of Research Classification (FoR), a part of the 2020 Australian and New Zealand Standard Research Classification (ANZSRC) system[6] has three hierarchical-levels namely Divisions, Groups and Fields. While Division represents a broad subject area, groups and fields represent detailed subsets of these areas. The Division and Group level of FoR has been emulated by Dimensions based on a machine-learning approach. The implementation in Dimensions

is categorised in 2-digit and 4-digit codes in which records are classified into 22 Divisions and 171 Groups. For example, the common record among the three databases with DOI '10.1016/j. tourman.2023.104385' contained data as '4605 Data Management and Data Science; 46 Information and Computing Sciences'. This database also provides alignment with many national subject classification schemes too. Moreover, the Dimensions database also provides a classification/ tagging of publication records on sustainable development goals.

Classification of research articles into different subject areas is an extremely important task in bibliometric analysis and information retrieval. There are primarily two kinds of subject classification approaches used in different academic databases: journal-based (aka source-level) and article-based (aka publication-level). The two popular academic databases- Web of Science and Scopus- use journal-based subject classification scheme for articles, which assigns articles into a subject based on the subject category assigned to the journal in which they are published. On the other hand, the Dimensions database uses an article-based subject classification scheme that assigns the article to a subject category based on its contents. A previous study provides a good comparative analysis of the subject classification schemes of the three databases.[64]

The availability of subject classification metadata is very important for the purpose of finding subject area distribution of publications from an institution, subject area distribution of

Table 8: Metadata Fields for Open Access.

Metadata Field	Web of Science	Scopus	Dimensions
OA Category	OA	Open Access	Open Access

Table 9: Metadata Fields for Conference Details.

Metadata Field	Web of Science	Scopus	Dimensions
Conference Title	CT	Conference name	-
Conference Year	CY	Conference date	-
Conference Location	CL	Conference location	-
Conference Sponsor	SP	-	-
Conference Host	HO	-	-
Conference Code	-	Conference code	-

Remarks: No conference details metadata in Dimensions.

Table 10: Metadata Fields for Subject Categories.

Metadata Field	Web of Science	Scopus	Dimensions
Subject Category	WC	-	Fields of Research (ANZSRC 2020), RCDC Categories, HRCS HC Categories, HRCS RAC Categories, Cancer Types, CSO Categories, Units of Assessment
Research Area	SC	-	-
Sustainable Development Goals	-	-	Sustainable Development Goals

highly cited publications, subject area-wise differentiation of open access patterns etc. The Web of Science and Dimensions metadata can be used for such analysis, however, Scopus metadata cannot be used for this purpose. Similarly, if it is required to identify the major subject areas contributing to research in an emerging topic (such as Artificial Intelligence or Quantum Computing), the subject classification metadata will be required for the analysis. Therefore, metadata provided by data downloaded from Scopus database through UI route cannot be used for such a study.

Dimensions database provides the additional tagging of publication records into various SDGs (Sustainable Development Goals). This tagging can be used for identifying major subject areas that contribute to research in a given SDG. Similarly, a country-specific publication analysis on SDGs can also be done by using Dimensions data, and demonstrated in a recent study.[65]

PRACTICAL IMPLICATIONS

The analysis of metadata provided by the three databases have shown that some kinds of scientometric analysis is not possible with metadata provided by a database. The analysis has also shown that there is an ease of doing different scientometric analysis if metadata from a particular database is used. This section presents a summary of such analysis, focusing mainly on the practical implications.

Analysis of possibilities of scientometric analysis

- The language composition analysis of publication records in a dataset is not possible with Dimensions provided metadata. Therefore, if the language of publications is to be analysed, then either Web of Science or Scopus should be used.

- The publication date of articles is not provided in the Scopus metadata and therefore an analysis involving use of the publication date information (such as to measure the speed of accumulation of altmetric mentions or citations) cannot be done with Scopus. For such analysis, Web of Science metadata should be used.

- The percentage of retracted papers can be obtained only from Web of Science and Scopus and not from Dimensions as it does not have such type of information included. Therefore, Web of Science or Scopus should be used for retraction analysis.

- Both Scopus and Dimensions do not provide ORCID id information of authors. Therefore, any analysis involving some linkage with ORCID id should use only Web of Science data.

- The Dimensions database does not provide ISSN information in its metadata and therefore either Web of Science or Scopus data can be used for any analysis involving ISSN information.

- The analysis of usage of articles requires usage count information, which is provided by the Web of Science database only and not by Scopus or Dimensions. Therefore, only Web of Science data can be used for article usage analysis.

- Altmetric analysis of publication records can only be done with Dimensions database as it is the only database providing the altmetric attention score. The data provided by the Web of Science and Scopus databases cannot be used for this purpose.

- The Dimensions database does not provide author or index keywords in user interface-based search and hence any analysis involving keywords (such as identifying thematic patterns and trends, co-word analysis etc.) is not possible with Dimensions database, leaving the choice of using either Web of Science or Scopus database.

- The Dimensions database cannot be used for an analysis of conference publications as it does not provide metadata about conferences. The Web of Science or Scopus database can be used for the purpose, though Scopus should be a preferred choice as it provides information about conference date whereas Web of Science only provides year information only.

- The Scopus metadata cannot be used for analysis involving subject category differentiation as it does not provide any metadata on subject classification in the downloaded publication metadata. Either Web of Science or Dimensions database can be used for such analysis.

- Dimensions database provides the additional tagging of publication records into various SDGs. This tagging can be used for identifying major subject areas that contribute to research in each SDG. Similarly, a country-specific publication analysis on SDGs can also be done by using Dimensions data. This analysis at present is not possible with metadata provided by Web of Science or Scopus.

Analysis of ease of doing scientometric analysis

Dimensions database explicitly provides the country of the authors in a dedicated Research Org Country field, making

analysis of ICP (Internationally Collaborated Papers) patterns easier. Similarly, Dimensions also provides a direct field for country names associated with each publication record. This allows a direct computation of the number of internationally collaborated papers in any set of publication records. Thus, international collaboration patterns in a data (including proportion of bilateral and multilateral collaborated papers) can be more easily computed using Dimensions data.

Dimensions database also provides city and state of publication records. This allows a direct computation of city-wise or state-wise research output, including collaboration patterns in them. The use of Web of Science or Scopus for the purpose will need additional string processing to extract such information. Therefore, Dimensions provides an ease of analysis in this regard.

The Dimensions database also provides standardized organization names with unique GRID Ids and organization type. This information makes it easier to compute university-industry, university-government, university-facility research collaboration. The use of Web of Science and Scopus data for this requires additional effort.

For the purpose of gender-based analysis of research publication, in Web of Science and Scopus databases extra processing effort is required to obtain full names of author and country of affiliating organization from different metadata fields. While Dimensions has more structured metadata for directly obtaining the country name of the affiliating organization.

The average page length of articles in a data set can be more directly computed in Web of Science and Scopus as compared to Dimensions, which will need an extra subtraction operation.

Any analysis involving creation of a citation network cannot be done easily with Dimensions publication metadata as it does not include cited reference details in the publication metadata. A separate download for citation metadata is required, which must be then matched with publication metadata through DOI matching. Therefore, for analysis involving co-citations or bibliographic coupling, Web of Science or Scopus database may be a preferred choice, if coverage is not an issue.

The Web of Science provides a cited reference count field, which can be used to directly compute the average number of references in any dataset being explored, which is otherwise difficult in Scopus and not possible in Dimensions.

Dimensions provide a more structured metadata for funding details. Funding country and ID are two additional fields in Dimensions, which can be used for country details of funding agencies and resultant publications in different subject areas

CONCLUSION

The study presented an analysis of metadata fields provided by the three scholarly databases- Web of Science, Scopus and Dimensions through UI based access. These metadata fields are grouped into various categories and the metadata fields in each group are compared across the three databases. Thus, the study provides a detailed account of the major metadata fields in data obtained from UI based search in Web of Science, Scopus and Dimensions and analyses the differences in the metadata fields across databases (**RQ1**). The analysis shows that different databases have some variations in the metadata fields provided in a publication record download from user interface-based search. Therefore, the effect of presence or absence of a metadata field on the possibility or ease of doing scientometric analysis is undertaken and key findings are summarized under the section on practical implications (**RQ2**). The findings can be useful for scientometric researchers, practitioners, and managers of these databases in various ways.

LIMITATIONS AND FUTURE WORK

The study is first of its kind to have analysed the metadata of the three important scholarly databases and there lies its novelty. However, the study only compares the metadata fields obtained through user interface-based search from the three databases. Since the databases provide different ways of access (such as through user interface or APIs), the functionalities and capabilities of scientometric analysis may vary with different forms of access. For example, the Dimensions database provides a very rich metadata through API based download, which is not only more structured but also provides more possibilities and ease of doing scientometric analysis as compared to user interface-based data download. Therefore, further analysis can be done to compare the metadata fields obtained through API-based data download route. Further, such a comparison may include metadata fields from other scholarly databases. In this connection, one may also note that there are different tools and packages available for performing scientometric analysis (such as VOS viewer[1]), which help in performing different kinds of scientometric analysis. Therefore, the ease of doing scientometric analysis also depends on compatibility of a database with the available tools and software. This aspect can be explored in a future work.

Another important aspect that this study has not addressed is the completeness of metadata fields provided by the three scholarly databases. It has been found in earlier studies[35] that all metadata fields provided by a database are not always populated with data for a given set of publication records and that there are missing values. For example, when we did a query on topic "sentiment analysis" in the three databases, we found that the metadata field for 'DOI' in Web of Science was 96% populated, in Scopus it was 80% populated and in Dimensions it was 90% populated. Similarly, the metadata for obtaining author address was 99% populated in

Web of Science, 94% in Scopus and 61% in Dimensions. Similar differences were also observed in many other metadata fields like Publication Year, Abstract, Funding Details, Open Access details etc. Dimensions database provides separate metadata explicitly for City, State and Country of research organization but those fields are not often fully differently populated, as was observed in case of the query "sentiment analysis", where the city details was only 62% populated, state details was only 36% populated, and country details was only 61% populated. Thus, it may be the case that in theory a database provides a metadata field, but in practice a good part of that metadata field may have missing values. Therefore, in actual practice not only the availability of a metadata field in a database matters for scientometric analysis, but the population of the metadata field is equally important. More research can be done in this direction to understand the impact of not only the availability of metadata fields but also the availability of data in them on the possibilities and the ease of doing scientometric assessment exercises.

CONFLICT OF INTEREST

The authors declare that the manuscript complies with the ethical standards and there is no conflict of interest whatsoever.

PREVIOUS PRESENTATION

A previous version of this paper was presented at the 19th International Conference of the International Society for Scientometrics and Informetrics (ISSI) held during July 2-5, 2023 at Bloomington, Indiana, USA, and the proceedings version is available at: https://zenodo.org/records/8306017.

FUNDING

REFERENCES

1. Van Eck NJ, Waltman L. Citation-based clustering of publications using CitNetExplorer and VOSviewer. Scientometrics. 2017;111(2):1053-70. doi: 10.1007/s11192-017-2300-7, PMID 28490825.

2. Batagelj V, Mrvar A. Pajek-program for large network analysis. Connections. 1998;21(2):47-57.

3. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. ICWSM Third international AAAI conference on weblogs and social media. 2009;3(1):361-2. doi: 10.1609/icwsm.v3i1.13937.

4. Mongeon P, Paul-Hus A. The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics. 2016;106(1):213-28. doi: 10.1007/s11192-015-1765-5.

5. Harzing AW. Two new kids on the block: how do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? Scientometrics. 2019;120(1):341-9. doi: 10.1007/s11192-019-03114-y.

6. Martín-Martín A, Thelwall M, Orduna-Malea E, Delgado López-Cózar E. Microsoft academic, Scopus, dimensions. Scientometrics. 2021;126(1):871-906. doi: 10.1007/s11192-020-03690-4, PMID 32981987, Google Scholar.

7. Visser M, van Eck NJ, Waltman L. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. Quant Sci Stud. 2021;2(1):20-41. doi: 10.1162/qss_a_00112.

8. Singh VK, Singh P, Karmakar M, Leta J, Mayr P. The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. Scientometrics. 2021a;126(6):5113-42. doi: 10.1007/s11192-021-03948-5.

9. Singh P, Singh VK, Arora P, Bhattacharya S. India's rank and global share in scientific research: how data sourced from different databases can produce varying outcomes? J Sci Ind Res. 2021b;80(4):336-46.

10. Birkle C, Pendlebury DA, Schnell J, Adams J. Web of Science as a data source for research on scientific and scholarly activity. Quant Sci Stud. 2020;1(1):363-76. doi: 10.1162/qss_a_00018.

11. Baas J, Schotten M, Plume A, Côté G, Karimi R. Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quant Sci Stud. 2020;1(1):377-86. doi: 10.1162/qss_a_00019.

12. Herzog C, Hook D, Konkiel S. Dimensions: bringing down barriers between scientometricians and data. Quant Sci Stud. 2020;1(1):387-95. doi: 10.1162/qss_a_00020.

13. Mayr P, Walter AK. An exploratory study of Google Scholar. Online Inf Rev. 2007;31(6):814-30. doi: 10.1108/14684520710841784.

14. Gavel Y, Iselid L. Web of Science and Scopus: a journal title overlap study. Online Inf Rev. 2008;32(1):8-21. doi: 10.1108/14684520810865958.

15. López-Illescas C, de Moya-Anegón F, Moed HF. Coverage and citation impact of oncological journals in the Web of Science and Scopus. J Inf. 2008;2(4):304-16. doi: 10.1016/j.joi.2008.08.001.

16. López-Illescas C, de Moya Anegón F, Moed HF. Comparing bibliometric country-by-country rankings derived from the Web of Science and Scopus: the effect of poorly cited journals in oncology. J Inf Sci. 2009;35(2):244-56. doi: 10.1177/0165551508098603.

17. Vieira ES, Gomes JA. A comparison of Scopus and Web of Science for a typical university. Scientometrics. 2009;81(2):587-600. doi: 10.1007/s11192-009-2178-0.

18. Chadegani AA, Salehi H, Yunus MM, Farhadi H, Fooladi M, Farhadi M, *et al.* A comparison between two main academic literature collections: Web of Science and Scopus databases. Asian Soc Sci. 2013;9(5):18-26. doi: 10.5539/ass.v9n5p18.

19. Harzing AW, Alakangas S. Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. Scientometrics. 2016;106(2):787-804. doi: 10.1007/s11192-015-1798-9, Google Scholar.

20. AlRyalat SA, Malkawi LW, Momani SM. Comparing bibliometric analysis using PubMed, Scopus, and Web of Science Databases. J Vis Exp. 2019;(152):e58494. doi: 10.3791/58494, PMID 31710021.

21. Aksnes DW, Sivertsen G. A criteria-based assessment of the coverage of Scopus and Web of Science. J Data Inf Sci. 2019;4(1):1-21. doi: 10.2478/jdis-2019-0001.

22. Yang K, Meho LI. Citation analysis: a comparison of Google Scholar, Scopus, and Web of Science. Proc Am Soc Inf Sci Technol. 2006;43(1):1-15. doi: 10.1002/meet.1450430118.

23. Falagas ME, Pitsouni EI, Malietzis GA, Pappas G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. FASEB J. 2008;22(2):338-42. doi: 10.1096/fj.07-9492LSF, PMID 17884971.

24. Torres-Salinas D, Lopez-Cózar ED, Jiménez-Contreras E. Ranking of departments and researchers within a university using two different databases: Web of Science versus Scopus. Scientometrics. 2009;80(3):761-74. doi: 10.1007/s11192-008-2113-9.

25. Mingers J, Lipitakis EA. Counting the citations: A comparison of Web of Science and Google Scholar in the field of business and management. Scientometrics. 2010;85(2):613-25. doi: 10.1007/s11192-010-0270-0.

26. Adriaanse LS, Rensleigh C. Comparing Web of Science, Scopus and Google Scholar from an environmental sciences perspective. S Afr J Libr Inf Sci. 2011;77(2):169-78. doi: 10.7553/77-2-58.

27. Adriaanse LS, Rensleigh C. Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison. Electron Libr. 2013;31(6):727-44. doi: 10.1108/EL-12-2011-0174.

28. De Winter JC, Zadpoor AA, Dodou D. The expansion of Google Scholar versus Web of Science: a longitudinal study. Scientometrics. 2014;98(2):1547-65. doi: 10.1007/s11192-013-1089-2.

29. Martín-Martín A, Orduna-Malea E, Thelwall M, Delgado López-Cózar ED. Google Scholar, Web of Science, and Scopus: a systematic comparison of citations in 252 subject categories. J Inf. 2018;12(4):1160-77. doi: 10.1016/j.joi.2018.09.002, Google Scholar.

30. Martín-Martín A, Orduna-Malea E, Delgado López-Cózar ED. Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison. Scientometrics. 2018;116(3):2175-88. doi: 10.1007/s11192-018-2820-9.

31. Thelwall M. Dimensions: A competitor to Scopus and the Web of Science? J Inf. 2018a;12(2):430-5. doi: 10.1016/j.joi.2018.03.006.

32. Huang CK, Neylon C, Brookes-Kenworthy C, Hosking R, Montgomery L, Wilson K, *et al.* Comparison of bibliographic data sources: implications for the robustness of university rankings. Quant Sci Stud. 2020;1(2):1-34. doi: 10.1162/qss_a_00031.

33. Visser M, van Eck NJ, Waltman L. Large-scale comparison of bibliographic data sources: Web of Science, Scopus, Dimensions and Crossref. In: Proceedings of the 17th international conference on scientometrics and informetrics; 2019. p. 2358-69.

34. Singh VK, Piryani R, Srichandan SS. The case of significant variations in gold–green and black open access: evidence from Indian research output. Scientometrics. 2020a;124(1):515-31. doi: 10.1007/s11192-020-03472-y.

35. Maddi A, Baudoin L. The quality of the web of science data: a longitudinal study on the completeness of authors-addresses links. Scientometrics. 2022;127(11):6279-92. doi: 10.1007/s11192-022-04525-0.

36. Andersen JP. Field-level differences in paper and author characteristics across all fields of science in Web of Science, 2000-2020. Quant Sci Stud. 2023;4(2):394-422. doi: 10.1162/qss_a_00246.

37. Zahedi Z, Haustein S. On the relationships between bibliographic characteristics of scientific documents and citation and Mendeley readership counts: A large-scale analysis of Web of Science publications. J Inf. 2018;12(1):191-202. doi: 10.1016/j.joi .2017.12.005.

38. Delgado-López-Cózar E, Martín-Martín A. Thomson Reuters uses altmetrics: usage counts for articles indexed in the Web of Science. Anu ThinkEPI. 2016;10:209-21. doi: 10.3145/thinkepi.2016.43.

39. Wang X, Fang Z, Sun X. Usage patterns of scholarly articles on Web of Science: a study on Web of Science usage count. Scientometrics. 2016;109(2):917-26. doi: 10.1007/ s11192-016-2093-0.

40. Liang G, Hou H, Hu Z, Huang F, Wang Y, Zhang S. Usage count: A new indicator to detect research fronts. J Data Inf Sci. 2017;2(1):89-104. doi: 10.1515/jdis-2017-0005.

41. Banshal SK, Singh VK, Muhuri PK. Can altmetric mentions predict later citations? A test of validity on data from ResearchGate and three social media platforms. Online Inf Rev. 2021;45(3):517-36. doi: 10.1108/OIR-11-2019-0364.

42. Costas R, Zahedi Z, Wouters P. Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. J Assoc Inf Sci Technol. 2015;66(10):2003-19. doi: 10.1002/asi.23309.

43. Haustein S, Peters I, Bar-Ilan J, Priem J, Shema H, Terliesner J. Coverage and adoption of altmetrics sources in the bibliometric community. Scientometrics. 2014;101(2):1145-63. doi: 10.1007/s11192-013-1221-3.

44. Thelwall M. Early Mendeley readers correlate with later citation counts. Scientometrics. 2018b;115(3):1231-40. doi: 10.1007/s11192-018-2715-9.

45. Thelwall M, Nevill T. Could scientists use Altmetric. com scores to predict longer term citation counts? J Inf. 2018;12(1):237-48. doi: 10.1016/j.joi.2018.01.008.

46. Banshal SK, Gupta S, Lathabai HH, Singh VK. Power Laws in altmetrics: an empirical analysis. J Inf. 2022;16(3):101309. doi: 10.1016/j.joi.2022.101309.

47. Seglen PO. Why the impact factor of journals should not be used for evaluating research. BMJ. 1997;314(7079):497. doi: 10.1136/bmj.314.7079.497.

48. Alberts B. Impact factor distortions. Science. 2013;340(6134):787. doi: 10.1126/scien ce.1240319, PMID 23687012.

49. Purkayastha A, Palmaro E, Falk-Krzesinski HJ, Baas J. Comparison of two article-level, field-independent citation metrics: Field-Weighted Citation Impact (FWCI) and Relative Citation Ratio (RCR). J Inf. 2019;13(2):635-42. doi: 10.1016/j.joi.2019.03.012.

50. Waltman L, van Eck NJ. Field-normalized citation impact indicators and the choice of an appropriate counting method. J Inf. 2015;9(4):872-94. doi: 10.1016/j.joi.2015. 08.001.

51. Janssens AC, Goodman M, Powell KR, Gwinn M. A critical evaluation of the algorithm behind the Relative Citation Ratio (RCR). PLOS Biol. 2017;15(10):e2002536. doi: 10.13 71/journal.pbio.2002536, PMID 28968388.

52. Bloudoff-Indelicato M. Text-mining block prompts online response. Nature. 2015;527(7579):413. doi: 10.1038/527413f.

53. Bornmann L, Haunschild R. Does evaluative scientometrics lose its main focus on scientific quality by the new orientation towards societal impact? Scientometrics. 2017;110(2):937-43. doi: 10.1007/s11192-016-2200-2, PMID 28239207.

54. Auranen O, Nieminen M. University research funding and publication performance—an international comparison. Res Policy. 2010;39(6):822-34. doi: 10.1016/j.respol.201 0.03.003.

55. Hicks D. Performance-based university research funding systems. Res Policy. 2012;41(2):251-61. doi: 10.1016/j.respol.2011.09.007.

56. Bloch C, Sørensen MP. The size of research funding: trends and implications. Sci Public Policy. 2015;42(1):30-43. doi: 10.1093/scipol/scu019.

57. Lathabai HH, Nandy A, Singh VK. x-index: identifying core competency and thematic research strengths of institutions using an NLP and network based ranking framework. Scientometrics. 2021;126(12):9557-83. doi: 10.1007/s11192-021-04188-3 .

58. Lathabai HH, Nandy A, Singh VK. Institutional collaboration recommendation: an expertise-based framework using NLP and network analysis. Expert Syst Appl. 2022;209:118317. doi: 10.1016/j.eswa.2022.118317.

59. Sun SL, Peng MW, Lee RP, Tan W. Institutional open access at home and outward internationalization. J World Bus. 2015;50(1):234-46. doi: 10.1016/j.jwb.2014.04.003.

60. Laakso M, Welling P, Bukvova H, Nyman L, Björk BC, Hedlund T. The development of open access journal publishing from 1993 to 2009. PLOS ONE. 2011;6(6):e20961. doi: 10.1371/journal.pone.0020961, PMID 21695139.

61. Chan L, Kirsop B, Costa SM, Arunachalam S. Improving access to research literature in developing countries: challenges and opportunities provided by Open Access. World Library and Information Congress, Oslo. 2005. Available from: https://archive.ifla.org/ IV/ifla71/papers/150e-Chan.pdf.

62. Piwowar H, Priem J, Larivière V, Alperin JP, Matthias L, Norlander B, *et al.* The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. PeerJ. 2018;6:e4375. doi: 10.7717/peerj.4375, PMID 29456894.

63. Srichandan SS, Piryani R, Singh VK, Bhattacharya S. The Status and Patterns of open access in Research Output of Most Productive Indian Institutions. Journal of Scientometric Research. 2020;9(2):96-110. doi: 10.5530/jscires.9.2.13.

64. Singh P, Piryani R, Singh VK, Pinto D. Revisiting subject classification in academic databases: A comparison of the classification accuracy of Web of Science, Scopus & Dimensions. J Intell Fuzzy Syst. 2020b;39(2):2471-6. doi: 10.3233/JIFS-179906.

65. Singh A, Kanaujia A, Singh VK. Research on Sustainable Development Goals: how has Indian Scientific Community Responded? J Sci Ind Res. 2022;81(11):1147-61.