Bulletin of the AAS

Research Data in Astronomy: Assessing its Impact and Visibility through a Bibliometric Analysis

Patricio Cortés-Rodríguez[1], Denise Depoortere[1], Lucy Opazo-Calfin[1]

[1]Bibliotecas, Pontificia Universidad Católica de Chile

Published on: Apr 27, 2022

ABSTRACT

In the context of the Open Access and Open Science initiatives, Research Data Management becomes relevant and of current interest to researchers, libraries, and institutions that contribute to the development of scientific knowledge. Particularly in Astronomy, it is a challenge to convert, standardize, process, and arrange large volumes of information and data (estimated in several petabytes), coming from the results of astronomical observations, astronomical catalogs, computational codes, among others, which constitute a fundamental initial step in scientific research [1]. Due to this, and the limited studies on this subject, a bibliometric methodology was applied to analyze indexed datasets in the Data Citation Index (Web of Science, Clarivate Analytics) in the Astronomy & Astrophysics category, for the period 2010 - 2019, allowing to identify annual evolutions, countries, and institutions with higher productivity, main repositories, and hosting platforms used in publications indexed in Web of Science. The analysis was complemented with InCites (Clarivate Analytics) bibliometric tools to determine the linkage of the publications with the datasets and their impact. The results from this study show a substantial increase in the number of indexed datasets provided by institutions in the United States, United Kingdom, Spain, and Chile, during the analyzed period. In turn, the publications that cite the main datasets show superior performances in the number of citations and the standardized impact, among other productivity and impact metrics.

1 Introduction

This study aims to contribute to the understanding of the need to share data in Astronomy [2]. Considering that this area has a long history of acquisition, systematization, and interpretation of large amounts of data [3], among which astronomy as a discipline generates a huge amount of unique and unrepeatable data [4] and is also a pioneer in open access to both publications and data.

Consequently, its preservation is also of interest, to facilitate that the data is made available and can be reused in other studies [5]. Therefore, and taking into consideration the above, one of the goals of this work is carrying out an analysis of the astronomy research data that have been cited and the metrics that can be obtained from them, using a tool that allows for such survey.

Therefore, this study focuses on astronomy research data in the world, its impact and visibility, knowing in which repositories they are deposited, and which are the countries that produce them the most, all the above, according to the Data Citation Index database (hereafter DCI), belonging to Web of Science (hereafter WoS) from the provider Clarivate Analytics.

To start this work, the research data was extracted from DCI that provides information from approximately 435 repositories from different disciplines.

Thus, under the area of "Physical Sciences", approximately 22 astronomy repositories were found, including NASA/IPAC, Mikulski, LAMBDA, and the Strasbourg CDS where astronomers deposit research data from areas such as cosmology, planets, and data Hubble Space Telescope, among others. These repositories are also used by researchers from the Astrophysics Institute of the Pontificia Universidad Católica de Chile.

2 Objectives

The main objective of this article is to carry out a bibliometric analysis of data sets indexed in Data Citation Index with Astronomy & Astrophysics category, for the period 2010 - 2019. In addition, the specific objectives of this work are: (1) Show the increase in indexed datasets, identifying the countries and institutions with the highest productivity; (2) Identify main repositories and hosting platforms; (3) Analyze the impact of datasets on publications indexed in Web of Science (Core Collection) - InCites.

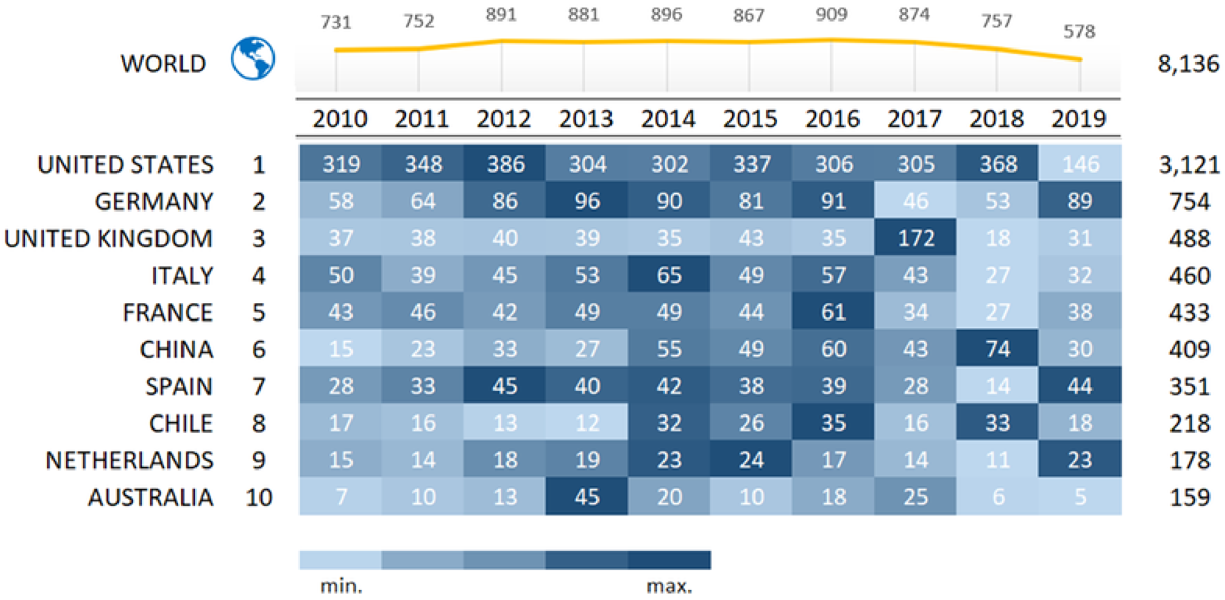
3 Methodology

The methodology used in this work is based on two consecutive processes. First, a systematic analysis was applied through the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) protocol, that enabled a replicable and transparent literature review to be carried out, which allowed for establishing inclusion and exclusion criteria for the datasets, in addition, sources of information, parameters, and search strategies were established. Thus, the datasets met the following criteria: a) they are published between 2010-2019, b) they are indexed in Data Citation Index, c) the affiliation of the first author is considered. Those records without information were completed through a manual compilation process, d) data in category WoS Astronomy & Astrophysics.

Secondly, the datasets were analyzed through a bibliometric methodology, using the Web of Science and InCites tools. These tools made it possible to identify the impact of the associated papers or those that cite the dataset. For these purposes, the following were considered: Times cited, accounted for the total number of citations received; Category Normalized Citation Impact (CNCI), allowed to evaluate the general impact of the papers in the category; Top 10%, indicated the percentage of high impact papers located within the 10% most cited; and, Quartile (Q), allowed to identify the position that the journal that published the paper occupies regarding the set of others in the same category, through its division into four groups (Q1 to Q4). Thus, the articles assigned in Q1 correspond to those located in the top 25% of the set of journals that presented the highest Impact Factor in 2019 (JIF).

4 Results and Analysis

The analysis found 8,136 Astronomy datasets indexed in DCI, for the analyzed period. According to Figure 1, between 2010 and 2016 an annual increase was observed that, globally, reached a maximum of 909 datasets deposited in 2016. As of 2017, it is possible to identify a decrease compared to the previously indicated period (2010-2016), which dropped to 578 datasets in 2019.



		2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	
WORLD		731	752	891	881	896	867	909	874	757	578	8,136
UNITED STATES	1	319	348	386	304	302	337	306	305	368	146	3,121
GERMANY	2	58	64	86	96	90	81	91	46	53	89	754
UNITED KINGDOM	3	37	38	40	39	35	43	35	172	18	31	488
ITALY	4	50	39	45	53	65	49	57	43	27	32	460
FRANCE	5	43	46	42	49	49	44	61	34	27	38	433
CHINA	6	15	23	33	27	55	49	60	43	74	30	409
SPAIN	7	28	33	45	40	42	38	39	28	14	44	351
CHILE	8	17	16	13	12	32	26	35	16	33	18	218
NETHERLANDS	9	15	14	18	19	23	24	17	14	11	23	178
AUSTRALIA	10	7	10	13	45	20	10	18	25	6	5	159

Figure 1. 2010-2019 annual evolution of dataset with Astronomy & Astrophysics category.

Figure 1 also displays this information disaggregated by country. In this way, a ranking of the countries with the largest number of deposited datasets was obtained. Top of the list are United States, Germany, and United Kingdom with a total of 3,121, 754, and 488 datasets, respectively. It is interesting to highlight Chile for being the only country in South America that presented a greater number of datasets, in line with its important contribution to scientific research in Astronomy [6].

Additionally, it is possible to observe the detail of the annual contribution, considering a color scale from minimum to maximum in terms of the number of datasets, for each country. In fact, a subperiod of higher productivity was confirmed between 2014 and 2016. However, after that date, a set of specific years with high productivity was observed. For example, 2018 registered the maximum values for United States, China, and Chile, or in 2019, for Germany, Spain, and the Netherlands. Therefore, the global decrease that was observed as of 2017 can be explained by the situation of the countries that are not part of the ranking in Figure 1.

Figure 2 shows a selection of institutions with the largest number of deposited datasets. Thus, it was observed that they belong to some of the 10 countries identified in the ranking of Figure 1. In this case, it is possible to highlight the institutions "NASA Goddard Space Flight Center" from United States, "Max Planck Society" from Germany, and “Istituto Nazionale Astrofisica (INAF)” from Italy, which are above the total average of datasets analyzed. Besides, high participation of US institutions was observed, at least 3 were included in this selection. UK and Germany obtained the presence of 2 institutions and again, Chile can be highlighted given its participation and leadership in research that was carried out in the “European Southern Observatory” in conjunction with Germany.

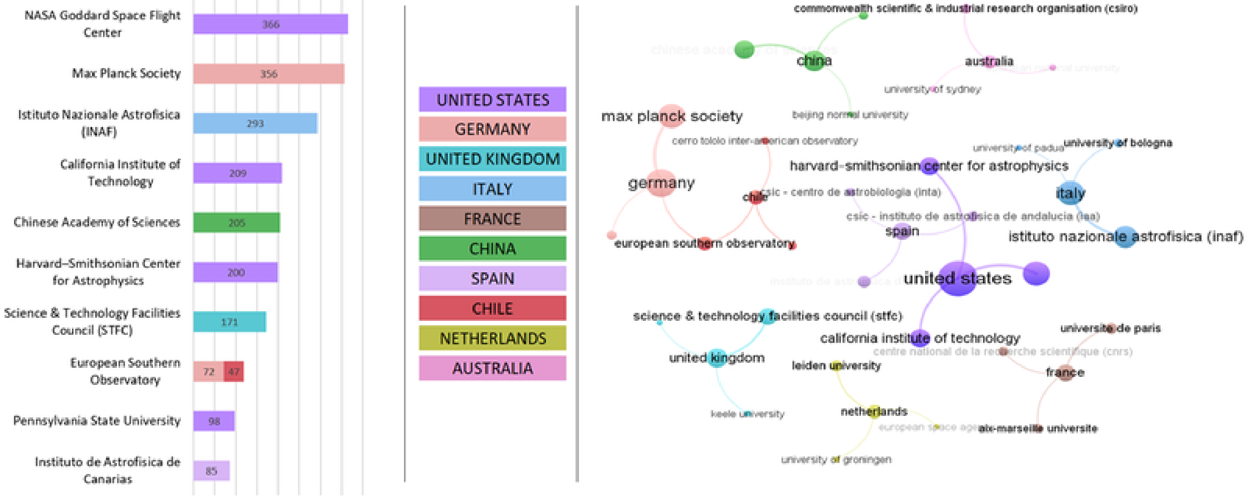


Figure 2. Selection of institutions with the largest number of datasets.

Additionally, Figure 2 presents a graph (located to the right of the image) number of datasets, which allows identifying a subset of institutions with higher productivity regarding to the prominent countries. In this way, 10 clusters were visualized that are referred to 10 countries and a selection of their institutions with the highest number of datasets. In addition, the size of the nodes made it possible to show the presence and importance of these institutions in the analysis carried out.

Another objective of this research was to identify the main repositories where the datasets were deposited. Thus, Figure 3 shows a selection of the most relevant repositories. It is possible to highlight the CDS repository in Strasbourg, France, since it was possible to observe that it included 90% of the total datasets analyzed. The UK eData repository and the USA LAMBDA continue to be important.

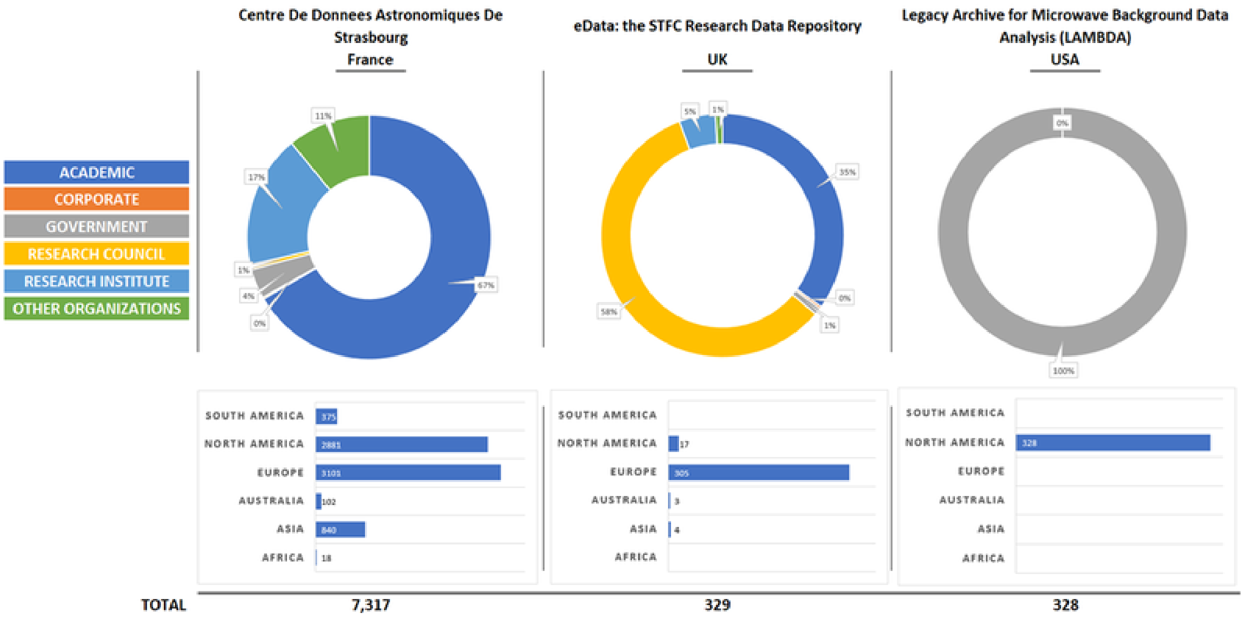


Figure 3. Analysis of the main astronomy repositories.

When considering the contribution by type of institution and region of membership, Figure 3 shows that those of an academic type are the main repositories of datasets. In the case of the CDS repository, it comprised 67%, while the eData repository, corresponded to 35%. In the case of the latter, it is also possible to highlight the European Research Council type with 58%. Likewise, 100% of the North American Government-type institutions deposited in the LAMBDA repository. Finally, it is possible to confirm that the CDS repository is very relevant for the deposit of datasets from countries from South America, North America, Europe, Australia, Asia, and Africa.

Finally, a set of bibliometric indicators made it possible to account for the impact of the papers indexed in Web of Science (WoS) which include at least one dataset. On the other hand, it was possible to identify the impact received by the new WoS papers that reused the datasets.

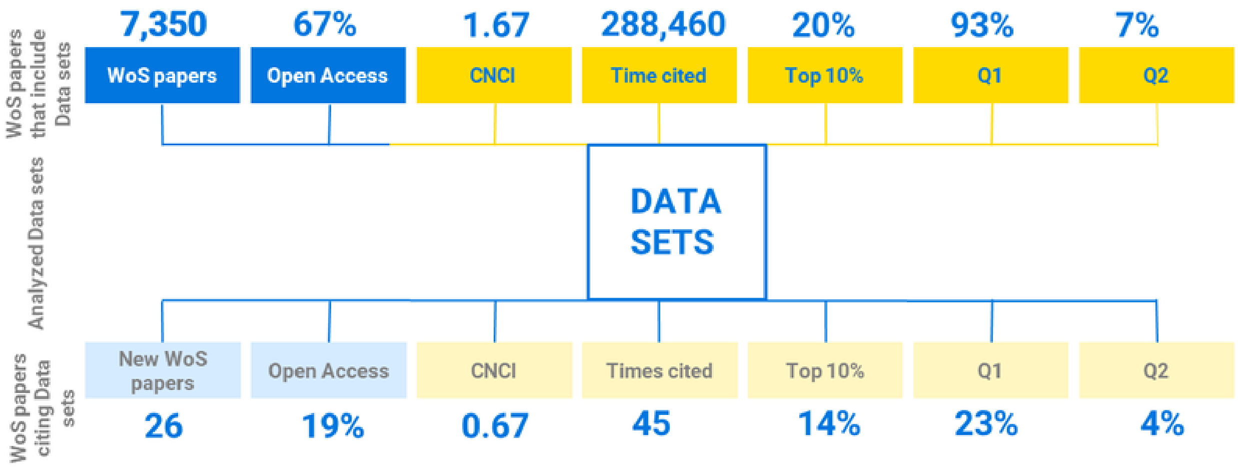


Figure 4. Impact of indexed publications in Web of Science that include dataset.

Figure 4 shows a set of metrics for the 7,350 WoS papers that gave rise to a subset of the database, corresponding to 7,432 records, the rest (704) were not associated with publications indexed in WoS. Regarding the WoS papers, it was identified that 67% are in open access (OA). These presented an outstanding Category Normalized Citation Impact (CNCI) above the value 1, placing it above the world average of impact for papers of this type. This was confirmed by observing other relevant impact indicators, for example the group received over 280 thousand citations; 20% were considered in the top 10% of the most select in scientific literature indexed in WoS; and, finally, 93% were published in high-impact journals, located in quartile 1 or 2.

On the other hand, 26 new WoS papers cited (reused) the analyzed datasets, of these, it was identified that 19% presented Open Access. In turn, it was possible to identify that the rest of the impact indicators were positioned below the expected values for the set of “WoS papers that include Data sets”.

5 Conclusions

This research has shown a sustained increase in datasets observed in the Astronomy & Astrophysics fields between 2010 and 2016. Standing out United States, Germany and United Kingdom as the countries that contributed the most by depositing their datasets in repositories included in DCI. It was identified that the CDS repository in Strasbourg in France contains over 90% of the analyzed datasets. Academic-type institutions being the ones that used this repository to a greater extent. Finally, a great impact of the publications that included at least one dataset was evidenced. 90% of the publications are indexed in WoS in high impact journals (Q1). However, low dataset citation was detected in new WoS publications (26 papers).

6 Data Availability

Underlying data

Repository UC: Datasets indexed in Data Citation Index in the Astronomy and Astrophysics category, 2010-2019. https://doi.org/10.7764/datasetuc/62181

This project contains the following underlying data:

dci_dataset_astronomy_2010-2019.xlsx (list of datasets exported from Data Citation Index)

7 Acknowledgments

The authors would like to thank the Prof. Wilfredo Palma for his help with the manuscript edition.

Citations

1. McCray, W.P. 2017, Osiris, 32, 243 doi:10.1086/707594 ↩

2. Zuiderwijk, A., & Spiers, H. 2019, International Journal of Information Management, 49, 228 doi:10.1016/j.ijinfomgt.2019.05.024 ↩

3. Brunner, R.J., Djorgovski, S.G., Prince, T.A., et al. 2002, Handbook of Massive Data Sets. Massive Computing, vol. 4 (Boston, MA: Springer) doi:10.1007/978-1-4615-0005-6_27 ↩

4. Pepe, A., Goodman, A., Muench, A., et al. 2014, PLoSO, 9, 104798 doi:10.1371/journal.pone.0104798 ↩

5. Borgman, C.L. 2017, ERCIM News. 100 https://ercim-news.ercim.eu/en100/special/if-data-sharing-is-the-answer-what-is-the-question ↩

6. Cortés, R., Depoortere, D., & Malaver, L. 2018, EPJWC, 108 doi:10.1051/epjconf/201818605002 ↩