Tirtharaj Dasgupta
Alumnus; Dept of Library and Information Science,
University of Kalyani
tirtharajdasgupta963@gmail.com

Facets of Openness of Research Spectrum of Chemistry in Asia: Data Carpentry-based Study

Tirtharaj Dasgupta

Abstract:

Openness has revolutionized the information society by loosening the limits on access to rich information. The benefits of openness have been felt more upon the availability of the Web. Open access, open data, open science, open research, etc. have been the talk of the time. Science has been able to move on from the typical "behind the laboratory" and "science only for the privileged" approaches to "open science" and "open research" approaches. It focuses on making different outputs of scientific research available openly. Such transparency has been able to significantly reduce the dominance of closed access science. Open science movement has made huge impact on the progress of modern society. Objectives: The objectives of the research is to study the various facets of openness, like open access status, open access colour and type of license used, of various research outputs in the domain of chemistry, produced from Asia in 2024, and to observe patterns in openness across the type, months and language of publications. Methodology: We have used bibliographic data from Open Alex, to study openness of the research outputs. We have taken a data carpentry-based approach using Open Refine. We have used the Unpaywall API service to study the different facets of openness using the DOIs. Findings: We observe that research in chemistry has mostly "not been open", The open outputs are mostly under CC BY license, and Gold open access. We also observe that in January, most of the publications have been done.

Keywords: *Chemistry, Data Carpentry, OpenAlex, Openness, OpenRefine, Open Access, Open Research, Open Science, Unpaywall*

1. Introduction:

The knowledge sphere is constantly growing. Researchers have been publishing their work in various forms like articles in journals, book chapters, etc. since ages. With the advent of the information and communication technologies, electronic journals and institutional repositories emerged, where e-resources of scholarly nature got published or archived. Scholarly world was dominated by resources that were behind the access wall. Most journals were published by commercial publishers, who charged huge amount of subscription fees, which automatically prevented many users from accessing the resources. This served as an unsuitable condition for the creators as well, who created the resources with their intellect, but they also required a subscription to access their own creation. Since knowledge is required to be accessed by all for the progress of the world, this called for a situation of open access to resources. With the

availability of the Web, creators are able to upload their articles in open access journals, so that they could be accessed by everyone free of all constraints of money, space and time. Open access has acted as a boon to the scholarly society. The openness in access isn't just limited to journal articles. Articles archived in repositories, various datasets, research reports, etc. are also open to access nowadays. This has led to the liberalization of access. Open access has been evolving with time, and it has been well backed by a strong set of licenses, that provide rules of access and usage of the resources. This has legalized open access alongside its liberalization, and also provided due rights to the creators as well.

2. Literature Review:

In a paper by Sheikh and Richardson, they focused on the progress in adoption of open access movement worldwide, with an emphasis on the role of libraries. They found that due to incurring of publication charges and the lack in availability of open access policies and mandates, the growth of open access in developing countries retarded as opposed to other nations (Sheikh & Richardson, 2023). Moore explored a different path in the history of open access publishing, where he looked into the contribution of researchers of humanities and social sciences in the progress of open access, before the onset of scientific open access publishing which is mostly considered to be the pushing force (Moore, 2020).

Heise and Pearce explored the complete potential of making scientific research in forms such as that of doctoral thesis, completely open, into the facet of open science, the comparison of publishing of physical science and humanities openly. It draws the conclusion that open publishing is a reality (Heise & Pearce, 2020). Mukhopadhyay designed a distributed weightage-based framework to measure open access friendliness of educational institutions (Mukhopadhyay, 2022). Another article defines various colours of open access publishing (*A Guide to Understanding the Colors of Open Access*, n.d.). Vrushali Sainath and Dhanmjaya looks into the progress in hospitality for open access in Western Asia, using various open access resource sites like DOAJ, OpenDOAR, etc. The lack in such progress in this area has been observed (Vrushali Sainath & Dhanmjaya, 2021).

A Wikipedia article provides a comprehensive note on the Creative Commons license, and various aspects related to it ("Creative Commons License," 2024). Barnabas et al. shows the extraction of structures related to chemical compounds from CORE and Google Patents full text document repositories. It uses open access cheminformatics toolkits, namely the Chemistry Development Kit (CDK), RDKit and OpenChemLib (OCL) to demonstrate the importance of structure normalization (Barnabas et al., 2022). Nelson and Eggett looks into the aspects of hybrid open access in the field of chemistry (Nelson & Eggett, 2017).

3. Research Gap:

The scholarly resources available on open access lack a proper metric study of the monthly patterns of open access colours and licenses. Moreover, a study on the open access pattern of research works in chemistry emanating from Asia in 2024 wasn't published.

4. Emergence of the Problem:

Chemistry, as a discipline, serves as the backbone of science. Constant research is required to discover more about the domain. Open access has served as a boon to the world of chemistry. Asia has been the provenance of many research works in this domain since ages, and a proper study of the pattern of open access publishing as well as the licensing pattern in 2024 would give an idea of the current status of evolution of open access.

5. Research Objectives:

(i) Check the patterns of OA colours and types of licenses of various research publications of different types in different months of 2024 (upto October 29, 2024), and study the pattern of open access publishing.

(ii) Design a scoring framework for measuring open access friendliness, and the licensing pattern of the resources.

(iii) Study the pattern of languages and resource types of open access publication of the resources in the year collectively.

6. Methodology:

This research is conducted with a data carpentry-based approach. We have used an open-source software OpenRefine for the work. This open-source application can be used to clean and transform data. We have used OpenAlex database to gather bibliographic data related to research publications in the field of Chemistry. It has a huge collection of bibliographic records related to publications produced globally. It provides various retrieval features. We export the bibliographic records as a single CSV file, and download it. This allows us to have all the bibliographic data in designated fields in a tabular form. We open a project in OpenRefine using the CSV file containing the bibliographic records. We can work with each column individually according to our requirements. We have the DOIs, that serve as unique identifiers representing the resources. To check the open access status of the research publications, we use the Unpaywall database, which is the largest bibliographic database for open access contents. It harvests open contents from publishers across the globe It allows free access to the dataset using a REST/API call with a call limit of 1,00,000 API calls per day. We make API calls based on the DOIs of the resources, in form of a GREL expression: https://api.unpaywall.org/v2/ + value + "email=tirtharajdasgupta963@gmail.com". This fetches the related OA data from the Unpaywall database in JSON format, that contains valuable data like information about open access nature

of the resource. We can extract relevant data from the JSON data using GREL expressions. We extract the months of publication, OA status and the OA colours. OA status indicates whether the resource is open access or not. The OA colour indicates the nature of open access. Lastly, we also extract the information related to the licenses adopted.

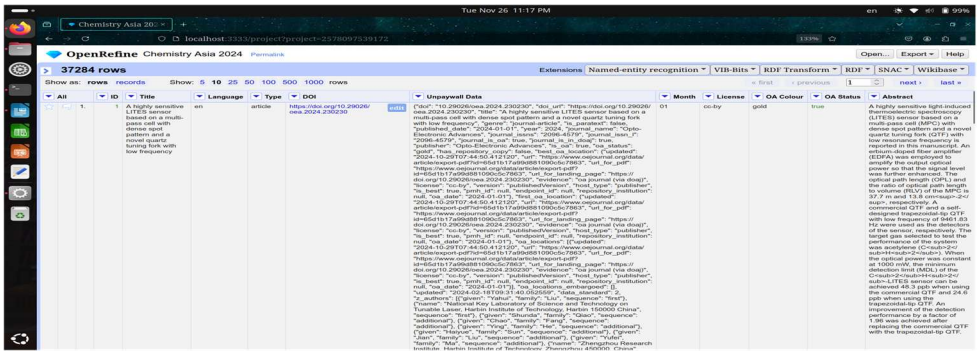

Figure 1: *The data of OA friendliness and licensing pattern across months, in OpenRefine*

We observe the results (from January 1, 2024 to October 29, 2024), and based on the OA colour, and kind of license adopted, we have designed a scoring system for each month, based on the various features of the open access colours and mostly the freedom provided by the licenses, as shown below:

Table 1: *Scoring system for the monthly observations of OA data, where Score = [0,1]*

Type of data	Value of the data	Score
OA Colour	Gold	1
	Green	0.75
	Bronze	0.5
	Hybrid	0.25
License used	CC BY	1
	CC BY-SA	0.75
	CC BY-NC	0.5
	CC BY-NC-SA	0.5
	CC BY-ND	0.5
	CC BY-NC-ND	0.25
	Others	0.5

Since some values are found to be blank, we did not consider the records of the resources for computation, to avoid complications. Upon assigning scores, we use the following method to get grand scores for each month, normalizing the grand scores: Grand OA Colour Score (GS_{colour}) = $\Sigma(S_{colour} \times N_{colour}) / N_{months}$, Grand License Score ($GS_{license}$) = $\Sigma(S_{license} \times N_{license}) / N_{month}$, and Average Grand Score ($GS_{average}$) = (GS_{colour} + $GS_{license}$) / 2 , where: S_{colour} = Score for a particular OA colour, N_{colour} = Number of publications for that particular OA colour in the month, $S_{license}$ = Score for a particular license, $N_{license}$ = Number of publications with a particular license in the month, N_{months}

= Total number of publications in that month, whose open access status data is available in the dataset (not blank), and N_{monthl} = Total number of publications in that month, whose licensing data is available in the dataset (not blank). These grand scores would represent the open access friendliness of research in that month.

7. Analysis of Data:

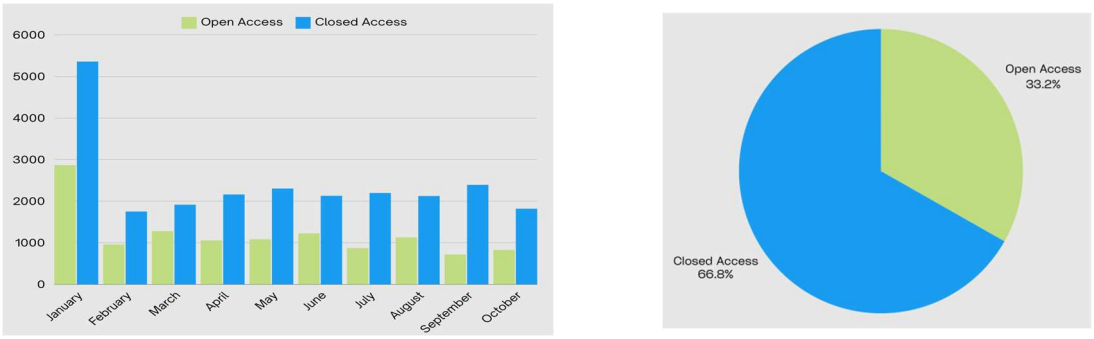

Figure 2. *OA status of the resources (upto October 29, 2024) – monthly and annual patterns*

From the above data, we observe that only about 33.21% of the total resources published in 2024 (up to October 29, 2024) are open access. The closed access resources mostly dominated till this date in this domain.

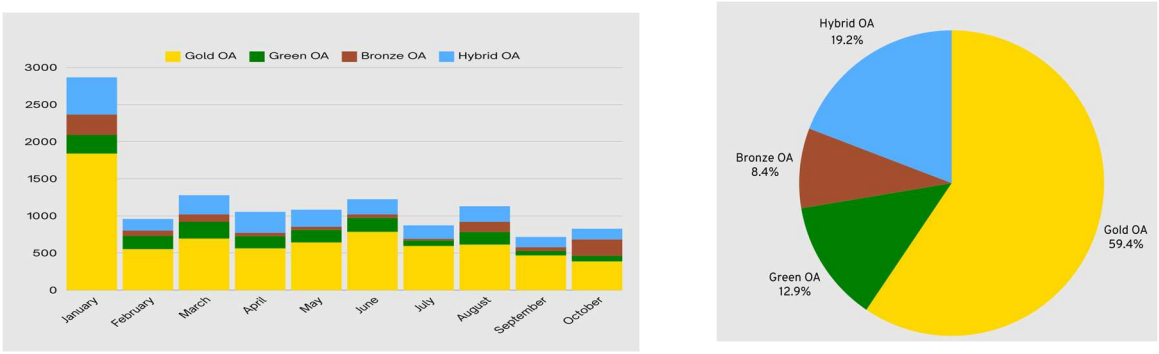

Figure 3. *OA colours of the resources (up to October 29, 2024) – monthly and annual patterns*

The dataset consisted of the 4 major open access colours: Gold, Green, Bronze, and Hybrid. We observe that mostly Gold OA resources dominate the group in each month, covering about 59.45% of the total open access resources in 2024.

About 60.95% of the total resources with license data available, have CC BY license, which provides the most freedom, requiring only attribution of the original creator, without any limitations like sharing alike, non-commercial work, and not producing derivative forms.

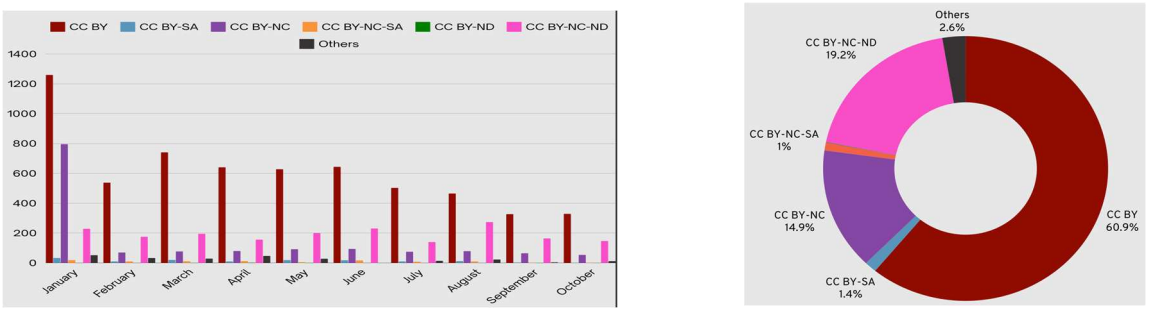

Figure 4. *Licenses of the resources (up to October 29, 2024) – monthly and annual patterns*

We assign grand scores to each month in a proper way, as follows:

Table 2. *Grand scores for all months (up to October 29, 2024)*

Month	GS_{colour}	$GS_{license}$	$GS_{average}$
January	0.278	0.743	0.511
February	0.281	0.772	0.530
March	0.306	0.804	0.600
April	0.242	0.800	0.521
May	0.250	0.776	0.513
June	0.298	0.762	0.530
July	0.230	0.792	0.511
August	0.265	0.694	0.500
September	0.184	0.718	0.451
October	0.223	0.740	0.482

We observe that in March, the average grand score is highest. This indicates that in this month, the scholarly publishing reached its best-case scenario, since the open access friendliness is the highest here, complemented with the liberalization provided by the licenses. The following illustrations indicates the patterns of languages and types of publications in the time period. Most of the publications were in English (about 98.69% of the OA publications whose language has been indicated), and most of the resources are articles (about 83.63%), followed by preprints (about 9.64%).

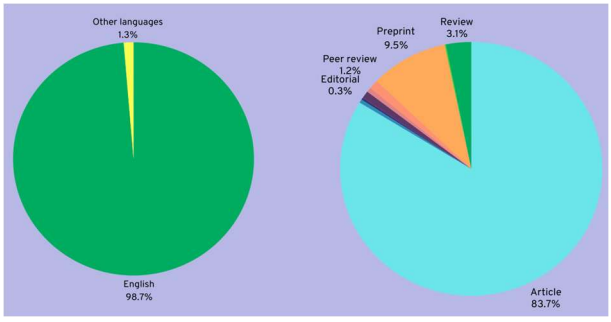

Figure 5. *Language and resource type of the OA publications (up to October 29, 2024)*

8. Results of the research:

The research findings mainly show that in the specified time period, closed access publications dominated the most (about 66.69%), as compared to the open access publications, which were just about 33.21%. Out of the open access publications, about 59.45% of them were Gold Open Access, which dominated the most. Upon studying the licensing pattern, we came to know that about 60.95% were associated with CC BY license, which was the most liberal license out of the ones used in our dataset. Most of the OA resources (about 98.69%) were in English. The open access resources were mostly in the form of articles (about 83.63%).

9. Drawbacks of the Study:

This research involved working with a dataset, which mostly consisted of data with no values assigned to them. Moreover, the dataset didn't provide a clear idea of exact date of publications in October.

10. Conclusion:

Open access is gradually adopted more across the world. Asia and the world as a whole, being a mixed bag of developed and developing nations, requires more R&D backing for its progress. Open access is hugely impacting the development by providing liberalization in accessing information resources. Still, more and more adoption is required, since the continent is still lacking enough OA resources emanating from it. The gradual adoption of openness will lead to a better and developed global society.

References:

A Guide to Understanding the Colors of Open Access. (n.d.). Retrieved November 25, 2024, from https://www.igi-global.com/newsroom/archive/guide-understanding-colors-open-access/4925/www.igi-global.com/newsroom/archive/guide-understanding-colors-open-access/4925/

Barnabas, S. J., Böhme, T., Boyer, S. K., Irmer, M., Ruttkies, C., Wetherbee, I., Kondić, T., Schymanski, E. L., & Weber, L. (2022). Extraction of chemical structures from literature and patent documents using open access chemistry toolkits: A case study with PFAS. *Digital Discovery*, *1*(4), 490–501. https://doi.org/10.1039/D2DD00019A

Creative Commons license. (2024). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Creative_Commons_license&oldid=1256987726

Heise, C., & Pearce, J. M. (2020). From Open Access to Open Science: The Path From Scientific Reality to Open Scientific Communication. *Sage Open*, *10*(2), 2158244020915900. https://doi.org/10.1177/2158244020915900

Moore, S. A. (2020). Revisiting "the 1990s debutante": Scholar-led publishing and the prehistory of the open access movement. *Journal of the Association for Information Science and Technology*, *71*(7), 856–866. https://doi.org/10.1002/asi.24306

Mukhopadhyay, P. (2022). *How green is my Valley? Measuring open access friendliness of Indian Institutes of Technology (IITs) through data carpentry* (Version Published (post-print)). Zenodo. https://doi.org/10.5281/ZENODO.6511080

Nelson, G. M., & Eggett, D. L. (2017). Citations, mandates, and money: Author motivations to publish in chemistry hybrid open access journals. *Journal of the Association for Information Science and Technology*, *68*(10), 2501–2510. https://doi.org/10.1002/asi.23897

Sheikh, A., & Richardson, J. (2023). Open access movement in the scholarly world: Pathways for libraries in developing countries. *Journal of Information Science*, 01655515231202758. https://doi.org/10.1177/01655515231202758

Vrushali Sainath, D., & Dhanmjaya, M. (2021). Open Access Initiatives in Western Asia. *Slavic & East European Information Resources*, *22*(1), 85–103. https://doi.org/10.1080/15228886.2021.1874349