Internationalization and Thematic Diversity in Data Use Within Open Research Infrastructures

A Scientometric Analysis of U.S. Department of Energy User Facilities

Lu Dong, Ren Wei, Yizhan Li, and Zexia Li
National Science Library
Chinese Academy of Sciences
Beijing, China
e-mail: donglu@mail.las.ac.cn, weir@mail.las.ac.cn, liyz@mail.las.ac.cn, lizexia@mail.las.ac.cn (corresponding author)

***Abstract*—The paper addresses limited empirical evidence on how data generated by large-scale research infrastructures are used in data-intensive and artificial intelligence-driven scientific research. The study is relevant to the conference theme as it examines data use patterns within open research infrastructures, contributing to the understanding of data-intensive science and infrastructure-based research systems. The paper presents an empirical evaluation based on micro-level administrative records of research proposals from United States Department of Energy user facilities over the period from fiscal years 2015 to 2025. It analyzes international participation, country-level distribution, thematic diversity, and temporal structural dynamics using a set of indicators that reflect participation scale, distributional diversity, and structural concentration, alongside entropy-based measures derived from project titles. The results show that international participation remained high, with a temporary decline during the COVID-19 pandemic followed by a rapid recovery, while thematic diversity remained consistently high and topic concentration low throughout the period. Overall, open research infrastructures exhibit strong resilience and sustained multidomain diversity in data use despite external disruptions.**

Keywords-open science; research infrastructure data; scientific resource; internationalization; thematic diversity.

I. INTRODUCTION

The paradigm of scientific research is undergoing a profound transformation, evolving from the traditional linear "observation-hypothesis-verification" model to two new paradigms: the Fourth and Fifth Paradigms. The "Fourth Paradigm" establishes data-intensive scientific discovery as a foundational new research framework [1], marking a critical shift away from conventional research logic. Building on this data-driven transition, recent advances in Artificial Intelligence (AI) and machine learning have further spurred the emergence of the "Fifth Paradigm," also referred to as AI4Science, distinguished by algorithm-guided discovery processes and human-AI collaborative research workflows [2]. Together, these paradigms represent the core dynamics of contemporary scientific restructuring.

With the rise of data-intensive and AI-driven research paradigms, research infrastructures increasingly extend beyond experimental operations to provide structured data services. Curated datasets, interoperable archives, long-term stewardship, and remote access mechanisms transform facilities into integrated data platforms. This transition is particularly evident in the U.S. Department of Energy (DOE) user facility system, where data management and open sharing are embedded alongside experimental access, serving as a typical example of this infrastructure-to-data-platform transition.

The U.S. Department of Energy (DOE) defined major research infrastructures funded by the federal government that provide open and shared access to researchers from academia and industry as National User Facilities in 2012 [3]. As of Fiscal Year (FY) 2025, the DOE operated 28 user facilities, covering areas including Advanced Scientific Computing Research (ASCR), Basic Energy Sciences (BES), Biological and Environmental Research (BER), Fusion Energy Sciences (FES), High Energy Physics (HEP), and Nuclear Physics (NP). In 2015, the DOE initiated the construction of a user project/experiment database [4].

This study is based on administrative records of research proposals from DOE user facilities covering FY 2015 to FY 2025. (The U.S. federal fiscal year runs from October 1 of the previous calendar year to September 30 of the current year.) The dataset includes user participation, institutional affiliations, and project descriptions. The analysis focuses on a subset of facilities that provide open access and structured data services, selected based on the availability of project-level proposal records. The resulting sample includes ARM, DIII-D, FACET/FACET-II, CINT, Alcator C-Mod, and NSTX-U.

As research infrastructures evolve into data platforms, issues of openness, accessibility, and global reach of infrastructure-generated data have become increasingly important. While open data principles, such as the Findable, Accessible, Interoperable, and Reusable (FAIR) principles, have been widely promoted, empirical evidence on who uses infrastructure-generated data and how such data contribute to knowledge production remains limited.

This study addresses the following research question: how are open research infrastructure data utilized in terms of international participation and thematic diversity, and what structural patterns characterize their usage over time? To answer this question, micro-level proposal data are analyzed using scientometric indicators and entropy-based measures.

By shifting the analytical focus from research outputs to data usage structures, this paper contributes to understanding open data ecosystems and the role of large-scale research infrastructures as global data platforms.

The remainder of this paper is organized as follows. In Section 2, the analytical framework and measurement design are described. In Section 3, the empirical results are presented. In Section 4, the conclusions and future research directions are presented.

II. METHODOLOGY

The analysis focuses on the structural characteristics of data use within open research infrastructures, rather than downstream research outputs. Attention is given to patterns of data usage across countries and thematic domains. This study aims to reveal the underlying structural features of open data ecosystems [5]. Data use within open research infrastructures is conceptualized along three analytical dimensions: (1) Internationalization-the extent to which open research data are used globally; (2) Topic Diversity-the breadth of research domains supported by the data infrastructure; (3) Structural Dynamics-the stability or evolution of usage patterns over time. These dimensions together capture the global reach, thematic inclusiveness, and longitudinal transformation of open data infrastructures.

A. Internationalization Measurement

To quantify international participation in open data use, three complementary indicators are adopted [6].

1) Non-U.S. User Ratio

The Non-U.S. User Ratio measures the proportion of users affiliated with Non-U.S. institutions in year *t*. Higher values indicate stronger international participation and a broader global diffusion of DOE open data resources. It is defined as:

$$NonUSRatio_t = \frac{N_{NonUS,t}}{N_{Total,t}} \quad (1)$$

where $N_{NonUS,t}$ represents the number of users affiliated with Non-U.S. institutions in year *t*; $N_{Total,t}$ represents the total number of users accessing DOE open data resources in year *t*.

2) Country Shannon Entropy

The Country Shannon Entropy measures the diversity of country participation [7]. Higher values indicate more even distribution across countries, reflecting broader international engagement. It is defined as:

$$H_t = -\sum_{i=1}^{N} p_{i,t} \ln p_{i,t} \quad (2)$$

where $p_{i,t}$ is the share of users from country i in year t; N is the total number of countries represented. Entropy is calculated using natural logarithms and is not normalized.

3) Country Concentration

The Herfindahl-Hirschman Index (HHI) is employed to measure concentration. Higher values indicate stronger dominance by a small number of countries, while lower values suggest more distributed global use. Entropy and HHI are used jointly to provide a balanced view of diversity and concentration. HHI is defined as:

$$HHI_t = \sum_{i=1}^{N} p_{i,t}^{\,2} \quad (3)$$

where p_i represents the share of users from country *i* in year *t*.

B. Thematic Diversity Metrics

Thematic diversity is derived from the textual analysis of Project/Experiment Titles, which represent the research purposes supported by the data infrastructure.

1) Keyword Distribution

Project/Experiment Titles were preprocessed using a standardized pipeline, including lowercasing, removal of general and domain-specific stopwords, and lemmatization to normalize morphological variants. Duplicate keywords within the same title were removed to avoid artificial inflation of term frequency. To reduce noise from unstable and weakly informative terms, keywords with an annual frequency below three occurrences were excluded. This filtering follows common text-as-data practice, as low-frequency terms are typically weakly discriminative and increase sparsity and computational burden without substantially affecting topic representations [8]. This consideration is particularly relevant for project/experiment titles, which are short texts with limited contextual and co-occurrence information. Therefore, a minimum annual frequency threshold of three was adopted to retain stable thematic signals while suppressing idiosyncratic noise [9]. After filtering, the annual keyword set was constructed, and relative frequencies were computed for entropy and concentration measurements.

2) Topic Shannon Entropy

Topic Shannon Entropy is employed to measure thematic diversity. Higher entropy indicates greater thematic diversity in data-supported research.

$$H_t^{\text{topic}} = -\sum_{j=1}^{M} q_{j,t} \ln q_{j,t} \quad (4)$$

where $q_{j,t}$ is the share of keyword or topic j in year *t*; M is the total number of topic clusters.

C. Structural Change Index

To evaluate temporal dynamics, the Structural Change Index (SCI) is computed as follows:

$$SCI_t = \sum_{k} \left| p_{k,t} - p_{k,t-1} \right| \quad (5)$$

where $p_{k,t}$ represents the share of country or topic *k* in year *t*.

SCI measures the magnitude of year-to-year structural shifts. Values close to zero indicate stability, while larger values reflect significant structural transitions in data usage patterns [10].

III. RESULTS

This section presents the empirical results of the study, focusing on international participation patterns and thematic diversity in the use of open research infrastructure data, along with their structural evolution over time.

A. International Participation in DOE Data-Providing Research Infrastructures

Based on the full dataset covering FY 2015-FY 2025, 10,113 user records were analyzed to evaluate international participation in research infrastructures that provide open data services. Internationalization is evaluated using three complementary indicators: the Non-U.S. User Ratio, Country Shannon Entropy, and Country Concentration.

1) Non-U.S. Participation Trends

Non-U.S. participation remained substantial throughout the study period, despite moderate annual variation (TABLE I). The Non-U.S. user ratio varied between 0.392 in FY 2018 and 0.318 in FY 2020. Following a notable decline between FY 2019 and FY 2020, the ratio gradually recovered, reaching 0.376 in FY 2024 and 0.379 in FY 2025.

The structural contraction observed in FY 2020 temporally coincides with the global COVID-19 pandemic, suggesting that global mobility restrictions may have contributed to fluctuations in international participation. Nevertheless, the pronounced post-2020 recovery demonstrates that open data ecosystems possess strong structural resilience and adaptive capacity in response to external disruptions.

TABLE I. INTERNATIONAL PARTICIPATION OF OPEN RESEARCH INFRASTRUCTURE DATA USE (FY 2015-FY 2025)

Year	International Participation		Country Structure			
	Non-U.S. Users	Non-U.S. Ratio	Countries Total	Country Entropy	HHI	SCI Country
2015	289	0.378	33	1.702	0.401	-
2016	352	0.392	41	1.759	0.385	0.067
2017	291	0.371	35	1.718	0.408	0.077
2018	286	0.392	36	1.743	0.385	0.087
2019	314	0.346	34	1.591	0.440	0.086
2020	285	0.318	37	1.495	0.476	0.077
2021	359	0.357	35	1.627	0.426	0.089
2022	314	0.328	36	1.537	0.462	0.070
2023	365	0.368	30	1.613	0.415	0.083
2024	384	0.376	31	1.656	0.405	0.042
2025	439	0.379	37	1.674	0.402	0.052

2) Country Diversity and Concentration

Country Shannon entropy further elucidates the structural evolution of global participation. The entropy value peaked in FY 2016 (1.759) and fell to its minimum in FY 2020 (1.495). The HHI exhibits a consistent pattern. Country concentration reached its maximum in FY 2020 (0.476) and declined gradually thereafter, falling to 0.402 in FY 2025.

Collectively, entropy and HHI measures demonstrate that FY 2020 represents a temporary phase of increased structural concentration, followed by a gradual diversification trend through FY 2025.

3) Structural Stability

The SCI indicator quantifies the year-to-year redistribution of country shares. Overall, the country-level SCI values remained relatively modest. The most pronounced structural shift occurred between FY 2020 and FY 2021, reflecting a post-contraction re-balancing of global participation patterns.

By contrast, the SCI for FY 2024-FY 2025 (0.052) was comparatively low, suggesting that the international usage structure had stabilized into a relatively steady configuration by the end of the observation period.

B. Thematic Diversity of Data-Supported Research

Thematic diversity is evaluated using keyword distributions extracted from Project/Experiment Titles. Topic Shannon Entropy and Topic SCI are used to measure diversity and temporal reconfiguration.

1) Topic Entropy

Topic entropy rose sharply from 4.780 in FY 2015 to 6.263 in FY 2016 (TABLE II), which may partly reflect database expansion and improved metadata registration in the early construction phase. From FY 2016 onward, topic entropy remained persistently high, ranging between approximately 6.11 and 6.36. Notably, it reached the highest observed value of 6.362 in FY 2025, indicating that DOE open data infrastructures support an increasingly diverse array of research activities. This sustained high level of entropy demonstrates that data usage has not become concentrated within specialized domains, but has instead continued to diversify across thematic areas.

TABLE II. THEMATIC STRUCTURE OF OPEN RESEARCH INFRASTRUCTURE DATA USE (FY 2015 – FY 2025)

Year	Topic Structure		
	Keywords_n	Topic_entropy	SCI_topic
2015	237	4.780	-
2016	991	6.263	0.633
2017	1031	6.263	0.295
2018	1082	6.233	0.303
2019	1100	6.200	0.335
2020	1072	6.223	0.265
2021	1138	6.106	0.357
2022	1176	6.244	0.259
2023	1183	6.217	0.241
2024	1216	6.257	0.206
2025	1285	6.362	0.288

2) Topic Structural Change

Topic SCI reached its maximum during the FY 2015-FY 2016 transition, consistent with the expansion of available datasets. After FY 2016, SCI values declined, indicating a gradual stabilization of the thematic structure. In FY 2025, topic SCI showed a moderate increase (0.288), suggesting a renewed redistribution of thematic emphasis rather than structural stagnation. This change may reflect the emergence of new research areas or enhanced cross-domain integration.

In summary, combining international and thematic indicators reveals several structural characteristics of DOE open research infrastructure data use. First, international participation remains resilient. Although international diversity exhibited a temporary contraction around FY 2020, the system subsequently recovered and re-diversified by FY 2025, indicating adaptive capacity in response to external disruptions. Second, thematic breadth remains high, as reflected by sustained levels of topic entropy from FY 2016-FY 2025. This suggests that open infrastructure data support multi-domain research rather than a narrow disciplinary focus. Third, SCI values indicate a pattern of gradual structural evolution rather than abrupt change, reflecting continuous adjustment within the data ecosystem, as opposed to pronounced volatility or disruption. Finally, by FY 2024-FY 2025, both country-level SCI and entropy indicators converge toward stabilization, implying that the system has transitioned into a relatively mature and stable configuration.

IV. Conclusion and Future work

The results indicate that DOE user facilities increasingly function as globally embedded data ecosystems rather than solely as physical experimental infrastructures. The sustained Non-U.S. participation ratio (approximately 0.37-0.39 in recent years), along with the recovery of country entropy after FY 2020, suggests that open research infrastructures extend beyond domestic use and operate within a distributed international knowledge network. This transformation aligns with the broader shift of large-scale research infrastructures toward structured data platforms that support global reuse through curated datasets, digital access mechanisms, and standardized metadata.

Despite a temporary rise in structural concentration around FY 2020, likely associated with global disruptions, the subsequent decline in the HHI and recovery of entropy indicate a re-diversification of participation patterns between FY 2021 and FY 2025. This trend, together with moderate SCI values, suggests that structural changes occur through gradual redistribution rather than abrupt transformation, reflecting adaptive stability within the system.

In parallel, consistently high topic entropy indicates that DOE open data resources support a wide range of research domains. The coexistence of thematic diversity and structural stability suggests that expansion occurs within a coherent platform structure rather than through fragmentation. The moderate increase in topic SCI observed in FY 2025 further suggests emerging thematic recombination, highlighting the role of open infrastructures in facilitating cross-domain knowledge integration in data-intensive scientific research.

It should be noted that proposal-level data reflect declared research intent rather than realized downstream scientific outputs. Nevertheless, such data can serve as a meaningful demand-side proxy for infrastructure data use, as research proposals represent anticipated data needs, planned methodologies, and targeted research questions. In this sense, proposal records provide an early-stage characterization of knowledge production activities, although their linkage to final scientific outputs remains indirect and probabilistic.

Future work will integrate publication data and citation networks to establish explicit connections between research proposals and downstream scientific outputs. This integration is expected to support a more comprehensive assessment of how infrastructure-generated data contribute to knowledge production, research impact, knowledge recombination, and cross-domain diffusion. In addition, more refined topic modeling approaches and cross-facility comparisons will be incorporated to further examine thematic evolution and the role of open research infrastructures in data-intensive science.

References

[1] K. M. Tolle, D. S. W. Tansley, and A. J. Hey, “The fourth paradigm: data-intensive scientific discovery [point of view],” Proceedings of the IEEE, vol. 99, no. 8, pp. 1334–1337, 2011.

[2] X. Li, and Y. Guo, “Paradigm shifts from data-intensive science to robot scientists,” Science Bulletin, vol. 70, no. 1, pp. 14–18, 2025.

[3] U.S. Department of Energy. *Definition of a user facility*. [Online]. Available from: https://science.osti.gov/-/media/_/pdf/user-facilities/memoranda/Office_of_Science_User_Facility_Definition_Memo.pdf

[4] U.S. Department of Energy. *User projects / experiments database for the office of science user facilities*. [Online]. Available from: https://science.osti.gov/-/media/_/pdf/user-facilities/memoranda/Office_of_Science_User_Projects_Experiments_Database_Memo.pdf

[5] C. L. Borgman, “The conundrum of sharing research data,” Journal of the American Society for Information Science and Technology, vol. 63, no. 6, pp. 1059–1078, 2012.

[6] L. Leydesdorff, “The static and dynamic analysis of network data using information theory,” Social Networks, vol. 13, no. 4, pp. 301–345, 1991.

[7] A. Stirling, “A general framework for analyzing diversity in science, technology and society,” Journal of The Royal Society Interface, vol. 4, no. 15, pp. 707–719, 2007.

[8] D. Maier, A. Niekler, G. Wiedemann, and D. Stoltenberg, ″How document sampling and vocabulary pruning affect the results of topic models, ″ Computational Communication Research, vol. 2, no. 2, pp. 139–152, 2020.

[9] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," presented at the Proceedings of the 22nd international conference on World Wide Web, Rio de Janeiro, Brazil, 2013. [Online]. Available from: https://doi.org/10.1145/2488388.2488514.

[10] C. S. Wagner, H. W. Park, and L. Leydesdorff, "The Continuing Growth of Global Cooperation Networks in Research: A Conundrum for National Governments," PLOS ONE, vol. 10, no. 7, p. e0131816, 201.