https://doi.org/10.51408/issi2025_126

Catalytic Effect of Open Data Platforms: The Case of the Global Biodiversity Information Facility (GBIF)

Honami Numajiri[1], Michio Oguro[2], Takayuki Hayashi[3]

[1] *doc22053@grips.ac.jp*, [3] *ta-hayashi@ grips.ac.jp*
National Graduate Institute for Policy Studies, 7-22-1 Roppongi, Minato-ku, Tokyo (Japan)

[2] *mog@ffpri.affrc.go.jp*
Forestry and Forest Products Research Institute, 1 Matsunosato, Tsukuba, Ibaraki (Japan)

Abstract

In recent years, data platforms have become essential to scientific research, and the Global Biodiversity Information Facility (GBIF) has emerged as a key infrastructure. However, how such platforms influence research trajectories remains poorly understood. This study examined GBIF's impact on researchers' topic selection by analysing papers before and after GBIF use and comparing them with non-GBIF users. Analysis of the research papers identified 20 distinct topics, revealing shifts in the research focus. GBIF users showed transitions from studies focused on ecosystem processes to research connecting ecosystem processes with plant-level characteristics, while maintaining unique patterns in forest biodiversity conservation. Furthermore, comparative analysis with non-GBIF users demonstrated that GBIF users exhibited unique topic transition patterns, particularly in areas such as plant-host interactions and habitat management, which indicate researchers' growing interest in this field, potentially aligning with increasing policy attention to biosecurity issues. These findings demonstrate GBIF's role as a catalyst in biodiversity science, actively shaping research directions rather than merely serving as a data repository. While highlighting GBIF's significant impact on scientific research, this study also identifies areas that require enhanced data coverage and accessibility to maximise the platform's scientific and societal contribution.

Introduction: The Evolution and Impact of Open Data in Scientific Research

The landscape of scientific research has been significantly transformed through open data policies, reshaping research practices, and knowledge dissemination. Government agencies, research institutions, and funding organizations actively promote research data release as a driver of scientific progress (Hrynaszkiewicz & Cadwallader, 2021). UNESCO defines Open Science as that which "enables free access to and reuse of scientific knowledge, thereby supporting collaboration and knowledge sharing" (UNESCO, 2021). Within Open Science, open research data is freely accessible, reusable, and properly documented data for scientific inquiry (OECD, 2015). Previous research has shown that open data policies increase citation rates, enhance collaboration opportunities, and improve research efficiency (Piwowar et al., 2007; Tenopir et al., 2011). Their significance lies in generating novel research through data reuse by diverse researchers (Borgman, 2015), particularly relevant as data-driven approaches become prevalent (Numajiri & Hayashi, 2024).

Two main types of Open Data exist: researcher-published data and Open Government Data (OGD). Researcher-published data comes from specific studies, providing insights into particular phenomena and enabling replication studies (Zuiderwijk et al., 2020). OGD, made available by government agencies, offers long-term datasets valuable for studying trends like climate change or population dynamics (Wirtz et al., 2022; Zuiderwijk & Janssen, 2014). Infrastructure challenges, particularly data fragmentation across institutions, remain a key barrier to efficient research data utilization (Quarati et al., 2021). While organizations like U.S. Geological Survey and various countries are developing centralized platforms to address this (Ojo et al., 2016), their effectiveness remains uncertain. Even when governments make data available, limited openness and continued fragmentation across agencies create barriers to data discovery and reuse (Wang & Shepherd, 2020).

GBIF: A Global Infrastructure for Biodiversity Open Data & Research question

The Global Biodiversity Information Facility (GBIF) [1] is a significant open data infrastructure in biodiversity research, providing worldwide access to biodiversity data. Before GBIF's 2001 establishment, accessing biodiversity information was challenging due to technical difficulties in data storage, insufficient training, and institutional sharing complications (Jones et al., 2006; Quarati et al., 2021). GBIF was designed to enable novel research through a single web interface (Telenius, 2011). Since 2012, it has seen explosive growth in scientific usage and currently hosts over 1.9 billion species occurrence records from thousands of institutions (GBIF, 2021). This extensive dataset supports applications from climate change assessment to species distribution modelling (Heberling et al., 2021). GBIF has enhanced its infrastructure through improved data management technologies and computational capabilities. Analysis of 4,000+ GBIF-enabled studies (2003-2019) shows species distribution modeling as a dominant application, though research applications have diversified (Heberling et al., 2021).

This study proposes viewing open data infrastructures like GBIF as enabling platforms in scientific research that facilitate novel combinations of research fields. Unlike temporary research funding, GBIF provides persistent infrastructure that enables researchers to integrate diverse data sources and explore new research directions. This catalytic effect leads researchers to discover and pursue new research topics they couldn't have explored before. The research question in this study is: **How does the adoption of GBIF data influence researchers' research topics?** The studies propose the following hypotheses: **GBIF data adoption leads to greater changes in research topics compared to non-users of GBIF data.** Through examining longitudinal changes in research content before and after researchers' initial use of GBIF data, this study investigates how open data infrastructure shapes the trajectory of scientific inquiry.

[1] https://www.gbif.org/

Data and Methodology

From the 'Peer-reviewed papers using data' section of the GBIF website, 10,915 papers with GBIF data DOIs were collected. Scopus search yielded 7,381 papers, from which 29,204 unique author IDs were extracted. All papers by these authors (762,123 papers) were collected through Scopus API. A control group was established from the top 20 journals that published most papers citing GBIF, out of 138 journals with 10+ papers citing GBIF. The top 20 journals yielded 46,811 non-GBIF authors who published a total of 1,156,791 papers. Changes in research topics were analyzed using five-year windows before and after authors' first GBIF use (reference year). Final analysis included 7,171 GBIF authors and 28,022 Non-GBIF authors with 3+ years of papers before/after reference years. The reference year was excluded to isolate GBIF's direct influence on subsequent research. For example, for a 2015 first use, analysis covered 2010-2014 and 2016-2020. For Non-GBIF Group authors, having no GBIF usage, each year from 2010 to 2020 was set as a potential reference year, creating 11 separate datasets per author.

For each author, titles and abstracts of their papers were combined into pre- and post-GBIF document sets. Titles and abstracts were preprocessed using NLTK and scikit-learn, converted to embeddings using Sentence Transformer, and clustered using K-means. Topic changes were measured using cosine distance between embedding vectors (0=identical, 1=dissimilar). To address sample size disparity (Non-GBIF Group having 11 datasets per author), bootstrap sampling matched case numbers between the groups by measuring GBIF Group author counts per reference year. Non-GBIF cases were randomly sampled 100 times to equalize author numbers for each reference year, enabling controlled analysis of topic changes.

Results1: Identification and Analysis of Research Topics

Based on evaluation metrics (silhouette=0.033, Calinski-Harabasz=7209.33, Davies-Bouldin=3.40) and elbow method analysis, the documents were clustered into 20 topics, striking a balance between interpretability and computational efficiency. Table 1 presents these 20 research topics with labels.

Table 1. Twenty research topics.

Topic	Topic label	Topic	Topic label
1	Forest carbon cycling models	11	Taxonomy & phylogeny
2	Conservation of avian habitats	12	Ecology of fish and fishery
3	Carbon stock in forests and biodiversity	13	Genomics of plant traits
4	Modelling	14	Limnology
5	Plant physiology of growth	15	Material cycling in soil
6	Plant diversity and traits in ecosystems	16	Effects of ecosystem changes on biodiversity and ecosystem services
7	Conservation of forest biodiversity and species	17	Forest pest management
8	Disease control	18	Water cycle model of forests
9	Population genetics, genetic diversity	19	Community of soil fungi and microbes
10	Conservation and management of marine ecosystem against climate change	20	Description of novel species

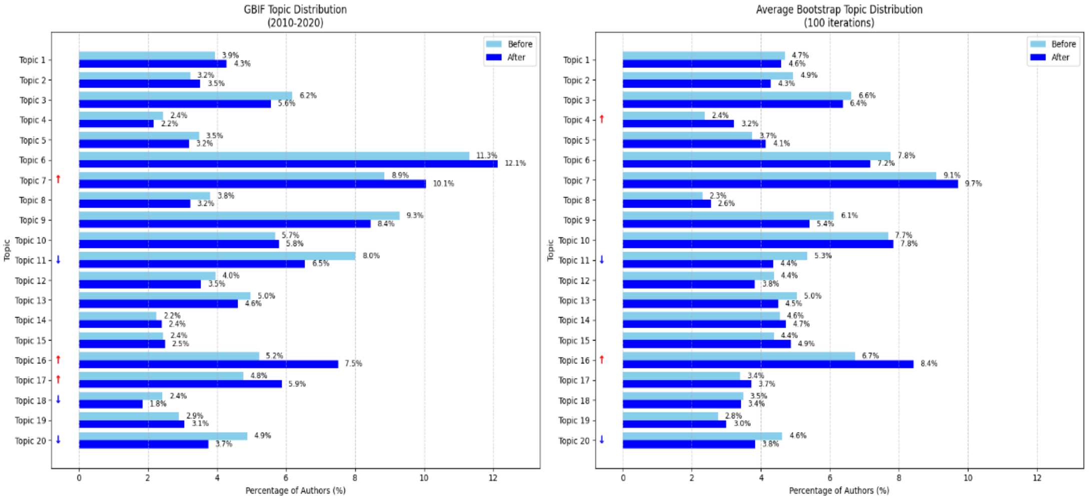

Figure 1 Topic distribution before and after GBIF use in GBIF and non-GBIF Groups (2010-2020).

Left: Topic distribution of GBIF group Right: Average topic distribution of bootstrap samples from non-GBIF group (100 iterations)

Note: The bars show the percentage of authors in each topic before (light blue) and after (dark blue) the reference year. For the GBIF group, the reference year corresponds to the year of their first publication utilizing GBIF data, while for the non-GBIF authors reference year were randomly sampled to match the number of GBIF authors who first used GBIF data in that year.

Comparing topic proportions before/after the reference year revealed key patterns. As shown in Figure , analysis of topic distributions revealed distinct research patterns between GBIF and non-GBIF groups. GBIF users maintained strong engagement in plant-focused research, with Topic 6 (Plant diversity and traits in ecosystems) showing the highest proportion (before: 11.3%, after: 12.1%). They also significantly increased their focus on forest biodiversity conservation (Topic 7: 8.9% to 10.1%, $p < 0.05$) and pest management (Topic 17). Both groups showed increased engagement in ecosystem services research (Topic 16: GBIF 5.2% → 7.5%, non-GBIF 6.7% → 8.4%, $p < 0.001$) and decreased focus on taxonomic studies (Topics 11 and 20). Significant differences in topic composition existed between groups both before ($\chi^2 = 523.400$, $p < 0.01$) and after ($\chi^2 = 626.012$, $p < 0.01$) the reference year. GBIF users showed consistently higher engagement in Topic 9 (Population genetics, genetic diversity), and lower in Topic 14 (Limnology) and 3 (Carbon stock in forests and biodiversity). Unique to the GBIF group were significant decrease in Topic 18 (Water cycle model of forests) ($p < 0.05$).

Results2: Transition Patterns

Given the differences in topic proportions between GBIF and non-GBIF groups described above, Figure 2 illustrates the differences in transition probabilities between topics, comparing GBIF and non-GBIF users. The analysis revealed several distinct patterns in research topic transitions. The most notable pattern centered on

Plant diversity and traits in ecosystems (Topic 6), where the GBIF group showed significant transitions from multiple topics. These included transitions from Topic 1: Forest carbon cycling models (GBIF group 17.1% vs. non-GBIF group 6.7%), from Topic 3: Carbon stock in forests and biodiversity (18.0% vs. 9.3%), and Topic 15: Material cycling in soil (17.3% vs. 11.3%). These transitions show that GBIF users shifted their research focus from forest-level studies to studies incorporating plant functional characteristics.

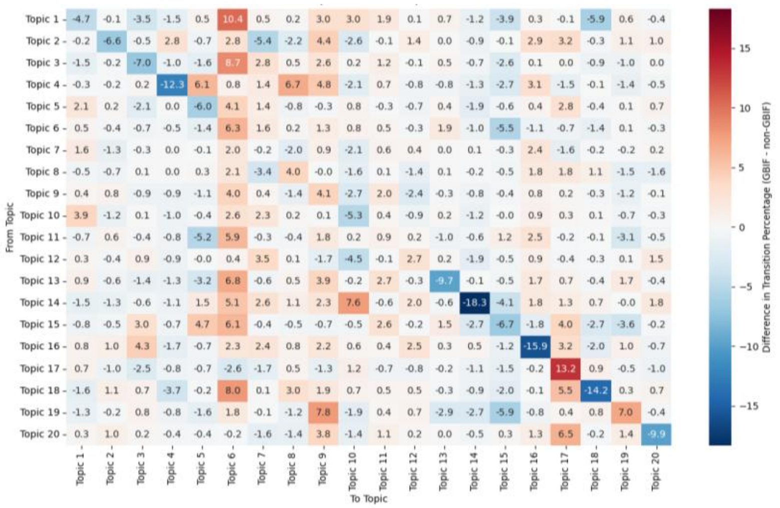

Figure 2 Comparison of topic transition patterns.

Note: Heatmap shows differences in transition probabilities between groups (percentage points). Red cells indicate higher probabilities in GBIF group, blue cells indicate higher in non-GBIF group. Values calculated as GBIF minus non-GBIF probabilities.

Results3: Quantitative Analysis of Topic Transition Distances

An analysis of topic transitions investigated whether GBIF users show greater transition distances between research topics compared to non-GBIF users. As shown in Figure 3, comparing cosine distances between the GBIF dataset (2010-2020, n=5,208, median=0.134, mean=0.135, SD=0.067) and 100 bootstrap samples (95% CI [0.122, 0.130]) revealed that GBIF users' transition patterns exceeded random expectations significantly ($p < 0.01$). The relatively small range of cosine distances (0.1-0.15) suggests that transitions represent expansions within related research domains rather than shifts to entirely different topics.

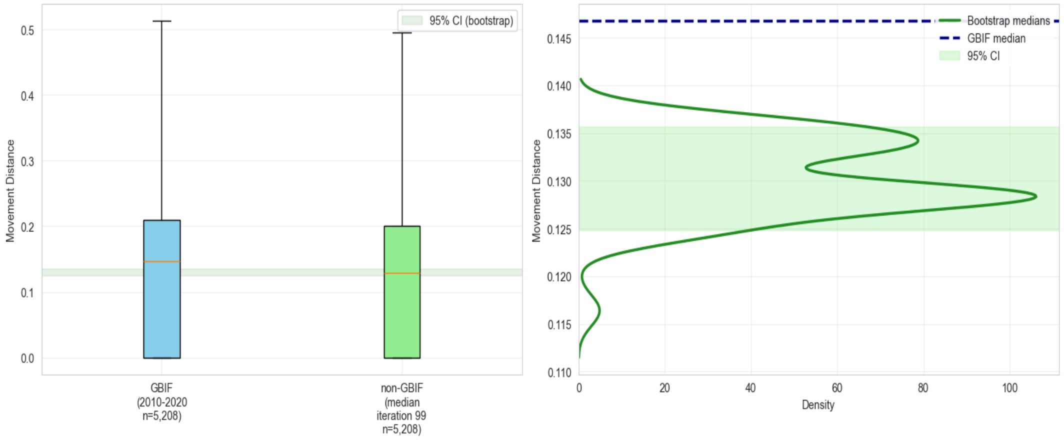

Figure 3 Comparison of Topic Transition Distances between GBIF Data and Bootstrap Samples.

Note: Left panel shows boxplots comparing GBIF and bootstrap samples distributions of non-GBIF. Right panel shows density plots of bootstrap medians, with GBIF median (dashed line) lying outside the bootstrap distribution's main density range (green shaded area).

Discussion and Conclusion

The study highlighted GBIF's impact on research trajectories through analysis of 20 distinct topics. The analysis suggests a possible relationship between GBIF use and researchers' transitions between ecosystem-level processes and plant-level characteristics. As noted by Mandeville et al. (2021), biodiversity platforms like GBIF enhance dataset accessibility, enabling researchers to bridge gaps between research domains. The finding that both GBIF and non-GBIF groups increased their engagement in ecosystem changes and biodiversity services research (Topic 16), while GBIF users showed distinct patterns in forest biodiversity conservation (Topic 7), suggests that while overall research trends may reflect broader scientific interests in the field, access to GBIF data might enable different approaches to biodiversity research.

Notably, the observed transition of GBIF users toward Topic 17 (Forest pest management) might indicate researchers' growing interest in this field. The observed transition patterns toward Topic 17 (Forest pest management) require careful interpretation, as the data suggests transitions from seemingly unrelated research areas such as species description (Topic 20), forest water cycles (Topic 18), and soil material cycling (Topic 15). These unexpected patterns indicate the need for more detailed investigation of how researchers actually shift between research topics.

The significantly higher transition distances in the GBIF group compared to non-GBIF users indicate that GBIF data facilitates broader research exploration, though within related domains as shown by the moderate range of cosine distances (0.1-0.15). This accessibility fosters diverse knowledge integration, leading to thematic diversification (Khan et al., 2021).

This empirical evidence revealed different patterns of research topic transitions between GBIF users and non-users, though the mechanisms and implications of these differences require further investigation.

Reference

Borgman, C. L. (2015). Big Data, Little Data, No Data: Scholarship in the Networked World. Big Data, Little Data, No Data. https://doi.org/10.7551/MITPRESS/9963.001.0001

Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B., & Schigel, D. (2021). Data integration enables global biodiversity synthesis. Proceedings of the National Academy of Sciences of the United States of America, 118(6), e2018093118. https://doi.org/10.1073/pnas.2018093118

Numajiri, H., & Hayashi, T. (2024). Analysis on open data as a foundation for data-driven research. Scientometrics, 1–18. https://doi.org/10.1007/S11192-024-04956-X

Ojo, A., Porwol, L., Waqar, M., Stasiewicz, A., Osagie, E., Hogan, M., Harney, O., & Zeleti, F. A. (2016). Realizing the innovation potentials from open data: Stakeholders' perspectives on the desired affordances of open data environment. IFIP Advances in Information and Communication Technology, 480, 48–59. https://doi.org/10.1007/978-3-319-45390-3_5

Telenius, A. (2011). Biodiversity information goes public: GBIF at your service. Nordic Journal of Botany, 29(3), 378–381. https://doi.org/10.1111/J.1756-1051.2011.01167.X

Vicent Civera, A., Baptista, P., Chatzivassiliou, E., Cubero, J., Cunniffe, N., et al. (2024). Commodity risk assessment of *Prunus cerasus* × *Prunus canescens* hybrid plants from Ukraine. EFSA Panel on Plant Health. *EFSA Journal*, 22 November 2024. https://doi.org/10.2903/j.efsa.2024.9089