１Ｄ２１

オープン研究情報基盤における日本語文献の可視性：人文学・社会科学評価における OpenAlex の可能性

○沼尻 保奈美(京都大学)、林 隆之 (政策研究大学院大学)

1. はじめに：オープン研究情報基盤としての OpenAlex

知識の公開と共有という原理は、科学の発展を特徴づける要素として伝統的に位置づけられてきた。マートン(Merton, 1973)が提唱した科学の行動規範「CUDOS」に示されるように、科学的発見の再現や知識の検証は、その前提として研究成果が公開され、共同体に共有されるという「共有主義(Communism)」によって支えられてきた。歴史的に見れば、15 世紀の活版印刷の発明、17 世紀以降の学術雑誌の創刊、19 世紀の学術団体の設立など、科学の進歩は常に知識共有の拡張と歩みを共にしてきた(Eisenstein, 1980; Stracke, 2020)。しかし、20 世紀後半以降、商業出版社による学術出版市場の寡占化により、本来開放的であった科学知識へのアクセスが制限される状況が生じた。近年のオープンサイエンスの理念(ユネスコ、2021 年)は、こうした閉鎖化への対抗として、科学本来の開放性を現代のデジタル技術によって再構築しようとする試みと位置づけられる。

こうした科学本来の開放性を現代において実現するための具体的な枠組みとして、2024 年 4 月にバルセロナ宣言(Barcelona Declaration on Open Research Information)が策定された[1]。すでに欧州を中心に 175 の研究機関がこの宣言への賛同を表明している(2025 年 9 月 27 日現在)。同宣言は、研究情報のオープン化を通じて、DORA や CoARA が進める研究評価システム改革を支える基盤の確立を目指すものであり、従来の商業データベースに依存した閉鎖的な評価体制からの転換を促している。バルセロナ宣言が提示する変革の方向性は多面的である。英語圏以外の研究成果の可視性向上、多様な地域・文化圏における学術的貢献の適切な評価、そして研究評価プロセス自体の透明性の確保といった、グローバルな学術コミュニティが直面する構造的課題への包括的な対応を目指している。

バルセロナ宣言が指摘するこれらの課題は、特に人文学・社会科学分野において深刻な形で現れている。現在の研究評価システムにおける言語的偏見の実態を見ると、Web of Science では英語が 93%、Scopus では英語が 88%を占めており、非英語圏の研究成果の可視性は極めて限定的である。Hicks(2004)が指摘するように、人文学・社会科学分野では書籍や地域的なジャーナルが重要な発表媒体となることが多く、研究対象の文化的・社会的文脈が研究成果の価値を決定する重要な要因となる。また、研究対象となる文化、歴史、社会現象の多くが特定の言語圏や地域に根ざしており、現地語による研究成果を排除することは、学術的多様性の著しい損失を意味する。ユネスコがオープンサイエンスの定義において「多言語の科学的知識」を明確に位置づけている(ユネスコ,2021 年)のは、こうした学問分野固有の特性や地域性・多様性の確保が科学的知識の包括性にとって不可欠であることを示している。

このような認識は、ヨーロッパにおける人文学・社会科学研究インフラの発展においても共有されている。2013 年 11 月にベルリンで開催された「未来に面する-人文社会科学のためのヨーロッパの研究インフラ」会議では、ヨーロッパ 19 カ国の専門家が集まり、人文学・社会科学分野における将来的課題として、学際的協力の促進と分野間の相乗効果の創出が重要な議題として議論された(ALLEA, 2013)。この会議で示された方向性は、単一の言語や地域に限定されない包括的な研究基盤の必要性を示唆している。

バルセロナ宣言が掲げる理念を体現し得るオープンな研究情報基盤の一つとして OpenAlex が国際的に期待されている。筆者らの先行的な分析によれば、OpenAlex に収録された論文(article)は、英語が 75%にとどまり、日本語が約 811 万件で英語に次ぐ第二の言語として収録されるなど、言語的多様性において顕著な違いを示している(沼尻ら,2025)。しかし、このようなマクロな集計を超えて、より詳細に、

[1] Barcelona Declaration on Open Research Information - Signatories. (2024). Retrieved March 26, 2025, from https://barcelona-declaration.org/signatories/

その多言語性や成果の多様性がどの程度包括的に担保されているのか、またデータの正確性や網羅性といった質的側面が十分に確保されているのかについては、未検証の課題である。

このような状況を踏まえ、本研究では以下の問いを設定する：

1. **OpenAlex における日本語文献の収録動向は時系列的に安定しており、研究評価の基盤として信頼性を有するか？：** 研究評価システムの基盤となるデータベースには、評価の一貫性の担保のため、収録データの時系列的安定性が求められる。急激な収録方針の変更や技術的問題による異常値は、評価結果の信頼性を損なう可能性があるためである。
2. **OpenAlex に収録された日本語文献の文献タイプ構成は、学術的多様性と質のバランスをどの程度実現しているか？：**人文学・社会科学分野では、Hicks(2004)が指摘するように書籍や地域的ジャーナルが重要な発表媒体となるため、論文のみならず多様な文献タイプの包含が学術的多様性の確保に不可欠である。
3. **OpenAlex の Concept 分析と Topic 分析を用いて、日本語文献(特に人文学・社会科学分野)はどのような研究領域と社会的課題との接点を示すか？：** OpenAlex ではこれまで「Concept」という Wikipedia 由来のテーマ分類が用いられてきており、公共的に共有される課題・テーマの語彙を含むため社会的課題との接点を検出しやすい。近年、Concept に変わり Topic という論文クラスタに基づく方法に変わり、それは研究実践の具体的展開を追跡できるはずである。この変更の時点を活用し、両者を併用することにより、日本の人文学・社会科学研究が地域的文脈と学術的価値をどのように結びつけているかを多角的に分析することが可能となる。

これらの問いを通じて、OpenAlex が収録する日本語文献の実態を明らかにし、バルセロナ宣言が目指す多言語的で包摂的な研究評価システムの実現に向け、OpenAlex が果たし得る役割を検討する。

2. データと分析

2.1 データ

本研究では、OurResearch が提供する OpenAlex データの 2025 年 3 月 31 日時点のデータダンプを使用した。OpenAlex は、学術論文、著者、機関、ジャーナル、研究トピック情報を含む包括的なメタデータを提供するオープンな学術データベースである。2025 年 3 月時点で約 1 億 9 千万件の学術出版物を収録している。同データベースは、Crossref、PubMed、arXiv、DOAJ 等の複数のソースから情報を統合し、DOI(Digital Object Identifier)を基軸とした文献同定システムを採用している。分析対象期間は、OpenAlex に収録された文献のうち、出版年が 2000 年から 2024 年の期間に該当する文献とした。日本語文献の特定については、言語フィールド(language)において「ja: 日本語」として分類された文献を対象とした。分野情報に関しては、従来の Concept と、より新しい Topic という二つの分類体系が提供されている。Concept 分類は、Wikipedia 由来の概念体系に基づいて文献に概念タグを付与するもので、階層的な構造(level 0 が最上位)を持ち、複数の概念が一つの文献に付与される。Topic 分類は、文献のタイトル、アブストラクト、出版元(ジャーナル)名、引用情報を考慮したアルゴリズムによって文献にトピックタグを付与するもので、約 4,500 のトピックが存在する[2]。各文献に対して各トピックのスコアが算出され、最も高いスコアを持つトピックがその文献の primary_topic として割り当てられる。Topic 分類の階層構造は、Domain→Field→Subfield→Topic の 4 層から構成される。本研究では、人文学・社会科学領域の特定と分析において、これら両方の分類体系を使用した。

2.2 分析手法

本研究では、以下の三つの分析手法を用いる。

第一に、研究問い 1 に対応する年別収録動向分析では、対象期間の日本語文献について各年の収録件数を算出し、その増減パターンを検証する。あわせて、データソースや収録ポリシー／処理仕様[3]の変更が件数に与える影響を点検し、研究評価基盤としての時系列的な安定性を検証する。

---

[2] https://docs.google.com/document/d/1bDopkhuGieQ4F8gGNj7sEc8WSE8mvLZS/edit#heading=h.5w2tb5fcg77r

[3] 旧 Microsoft Academic Graph からの継承・仕様更新の影響

第二に、研究問い 2 に対応する文献タイプ・出版社分析では、日本語文献を論文、書籍、会議録等の文献タイプ別に分類し、その構成比を算出する。また、出版社分析により収録内容の質的特徴を把握し、多様な発表媒体の包含状況を検証する。

第三に、研究問い 3 に対応する分野別分析では、Topic 分類(primary_topic)を用いて各文献の主トピックを特定し、そのトピックが属するドメインから人文・社会科学に該当する文献集合を抽出する。抽出後、(1) ドメイン別件数、(2) 上位トピックの頻度分布、(3) 文献タイプ構成、(4) 出版社分布を集計する。併せて、Concept 分類では level 2 の概念を対象に、社会課題関連語彙に合致する概念の出現件数を計測し、Topic による抽出結果とのカバレッジを比較する。最後に、Topic と Concept の結果を照合し、社会的課題との接点把握における両手法の有効性と限界を整理する。

3. 結果

3.1 言語別文献収録状況の概観

本分析で使用した OpenAlex データには、総計 196,332,212 件の文献が収録されており、55 の異なる言語が確認された。2000 年から 2024 年に出版され、言語情報が記録されている文献を対象とした言語別分析の結果(図 1)、英語文献が 139,069,836 件(74.6%)で最も多く、続いてスペイン語が 6,512,631 件(3.5%)で第二位、日本語が 5,338,183 件(2.9%)で第三位となった。これは筆者らの先行研究において article のみを対象とした分析で日本語が第二位であった結果とは異なる順位を示している。上位 10 言語の分布では、フランス語が 5,203,302 件(2.8%)で第四位、ドイツ語が 4,699,239 件(2.5%)で第五位、中国語が 4,042,581 件(2.2%)で第六位となった。ポルトガル語(2.0%)、韓国語(1.8%)、インドネシア語(1.5%)、ロシア語(1.2%)が続き、上位 10 言語で全収録文献の約 96%を占める。

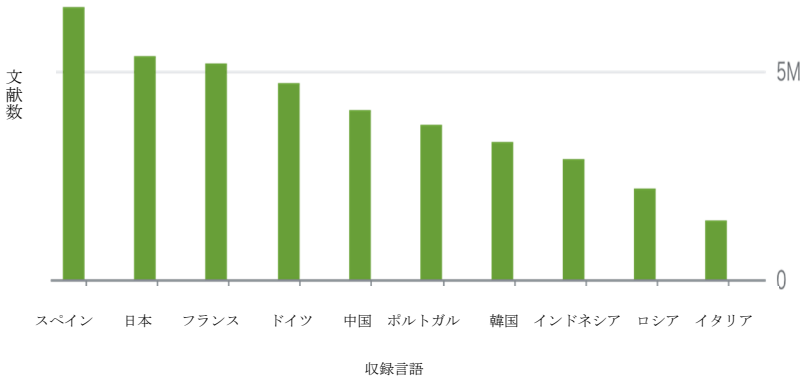

図 1 OpenAlex における言語別文献収録状況(上位 10 言語、2000–2024 年)

3.2 分析 1：日本語文献の時系列収録動向

日本語文献の年別収録件数を 2000 年から 2024 年まで分析した結果、明確な変動パターンが確認された(図 2)。日本語文献は 2000 年から 2015 年にかけて、収録件数は 211,196 件から段階的に増加し、2008 年に 294,979 件でピークに達した。その後、2015 年まで概ね 25 万件から 28 万件の範囲で推移し、比較的安定した収録動向を示した。2016 年に 396,373 件と前年比 1.6 倍の急増が観察されたが、翌 2017 年には 252,108 件へと大幅に減少した。2018 年以降は継続的な減少傾向が顕著となり、2018 年に 124,666 件と前年の半分以下に減少した。2019 年(140,366 件)、2020 年(151,893 件)と微増したものの、2015 年以前の水準には回復しなかった。2021 年以降の減少はさらに極端であり、80,061 件(2021 年)、11,592 件(2022 年)、12,021 件(2023 年)、6,067 件(2024 年)と推移し、2000 年代初期の水準を大幅に下回った。11,592 件(2022 年)、12,021 件(2023 年)、6,067 件(2024 年)と推移し、2000 年代初期の水準を大幅に下回った。全文献数は 2000 年の約 353 万件から 2020 年の約 1,105 万件まで概ね一貫した増加を示し、2021 年以降も 900 万件から 1,000 万件台で安定的に推

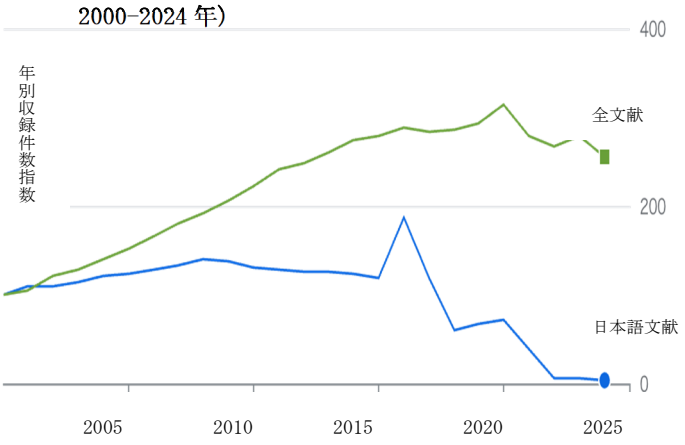

図 2 日本語文献と全文献の年別収録件数推移の比較
(2000 年から 2024 年、2000 年を 100 として指数化)

移している。日本語文献で観察された 2016 年の急増、2018 年以降の急減、特に 2022 年以降の極端な減少は、全文献の動向とは全く連動していない。

3.3 分析 2： 日本語文献の文献タイプ構成分析

日本語文献の文献タイプは、全期間を通じて article が 97.95%を占め、書籍は 1.18%、書籍章は 0.41%にとどまった。MAG 期間(2000-2021 年)では、article が 97.99%、書籍が 1.18%、書籍章が 0.41%を占めていた。MAG 除外後(2022-2024 年)では、article の割合が 92.27%に減少した。この減少分の主な要因は、OpenAlex の収録対象が拡大し、研究助成情報（grant）が 0.15%から 5.81%へと増加したことである。grant は学術成果物ではなく研究資金のメタデータであり、その増加は収録方針の変更を反映している。dataset(0.05%→1.05%)、review(0.07%→0.25%)、preprint(0.01%→0.25%)の割合も相対的に増加した一方で、書籍は 1.18%から 0.05%へ、書籍章は 0.41%から 0.16%へとさらに減少した。

出版社別では、Elsevier BV(14.51%)、Springer(3.40%)、Wiley(3.31%)など国際的商業出版社が上位を占めた。しかし、landing_page_url の分析により、これらに分類された日本語文献多くが機械翻訳メタデータであることがわかった。すなわち、MAG 期間に J-Global 経由で収録された日本語文献約 77 万件のうち、Elsevier が 87,141 件、Wiley が 17,558 件を占め、タイトルに「【JST・京大機械翻訳】」や「【Powered by NICT】」といった機械翻訳の表示が含まれている。また、文芸春秋(0.97%)、経済界(0.74%)、商業界(0.53%)など、一般向け出版社や商業誌も含まれていた。これらの文献は CiNii 由来の書籍章として収録されており、内容は学術論文ではなく、一般雑誌記事である（例：「東京五輪、国民は望むのか」「大谷翔平『二刀流』を開眼させた男」「お笑いコンビ かまいたち」等）。これは CiNii が収録対象とする雑誌記事が OpenAlex に流入し、学術文献以外のコンテンツが混入していたことを示している。MAG 除外前後での出版社構成の変化は大きく、文芸春秋(6,667 件)、経済界(5,067 件)、商業界(3,627 件)といった一般雑誌出版社は、MAG 除外後に完全に消失し、医学書院(15,298 件)や大学リポジトリ(広島大、筑波大、京都大)も MAG 除外後には収録されなくなった。対照的に、Elsevier、Springer、Wiley などの国際商業出版社や、日本の主要学会(循環器学会、心理学会、外科学会など)は収録が継続している。ただし、後者については前述の通り、J-Global 経由の機械翻訳メタデータが多く含まれる。

3.4 人文学・社会科学分野における日本語文献の研究領域の特徴

Topic 分類の primary_topic に基づいて分野を特定した結果、Social Sciences ドメインの合計は 21,484 件(全体の約 0.4%)に留まることが明らかになった。内訳は Social Sciences(狭義)が 10,355 件、Psychology が 4,128 件、Arts and Humanities が 2,266 件、Business, Management and Accounting が 2,279 件、Economics, Econometrics and Finance が 1,566 件である。Social Sciences ドメインの主要なトピックを確認した結果、Earthquake and Disaster Impact Studies(1,065 件)、Japanese History and Culture(882 件）、Urbanization and City Planning(732 件）、Urban Transport and Accessibility(614 件)、Asian Culture and Media Studies(563 件)など、災害対策、都市計画、地域文化といった日本社会に直接関連するテーマが上位を占める。Migration, Aging, and Tourism Studies(368 件)、Disaster Management and Resilience(291 件)といったトピックも確認され、日本が直面する高齢化や災害対策といった社会的課題と学術研究の接点がうかがえる。Arts and Humanities の 2,266 件についても詳細分析を行った結果、言語学関連のトピックが上位を占めている。Translation Studies and Practices(169 件)、EFL/ESL Teaching and Learning(151 件)、Language, Discourse, Communication Strategies(96 件)など、言語教育・翻訳研究が中心であり、伝統的な人文学の中核である文学(Literature and Literary Theory で最大 87 件)、歴史学(History で 29 件)、哲学(Philosophy で 42 件)の収録は極めて限定的である。

Concept 分析(level2)では、Politics(483 件)、China(186 件)、Tourism(139 件)など意味のある概念が確認される一方で、Process (computing)、Gene、Scale (ratio)、Current (fluid)、Work (physics)といった、他分野の用語が多数混入している。また、社会的課題との接点を検出する目的でキーワード検索を行った結果、Natural disaster(26 件)、Healthy aging(3 件)など、Topic 分析で確認された

Earthquake and Disaster Impact Studies(1,065 件)や Migration, Aging, and Tourism Studies(368 件)と比較して、はるかに少ない件数しか検出できなかった。

Social Sciences ドメイン全体における文献タイプ構成は、article が 89.2%、書籍が 8.85%であった。出版社からは、上位に都市計画学会(2,195 件)、土木学会(1,205 件)、日本心理学会(753 件)など日本の学会が中心を占めており、全日本語文献で上位であった国際商業出版社は含まれないという分野特有の傾向が確認された。

４．議論

本研究は OpenAlex における日本語文献の実態を三点から検証した。

第 1 に時系列的安定性については、研究生産の実変動というより収録プロセスの転換に起因する構造変化が明確に認められる。特に、MAG 由来レコードから DOI 等の国際識別子を基軸とする取り込みへ移行した時点を境に、系列の水準・構成が同質性を失っており、日本語文献の収録は従来の包括的収録から大幅に縮小され、日本国内で発行される学術文献の可視性が著しく低下した。

第 2 に文献タイプの多様性と質では、MAG 除外後に一般雑誌記事の混入は解消され質的担保は改善したが、書籍比率は 1.18%から 0.05%へ急減し、大学リポジトリや専門出版社の可視性も縮小した。人文・社会科学で重要な書籍・地域誌の包含という観点から、多様性の後退は看過できない。ただし、可視性の低下には、文芸春秋や経済界といった一般雑誌記事が除外されたことによるデータの質向上という側面も含まれており、収録件数の減少は必ずしも学術的価値の損失のみを意味するものではない。

第 3 に、社会的課題との接点の可視化では、Topic 分析により災害対策・都市計画・高齢化など日本社会に直結するテーマが抽出された。これに対し、Concept 分析は学際的一般語の付与が広く、社会科学固有の文脈を峻別しにくい（例：Process／Scale／Work が理工系概念として扱われやすい）。Topic はタイトル・抄録・誌名・引用情報を統合して解析できるため、文脈適合的な領域特定と課題抽出に相対的に優れる。ゆえに当初仮説に反し、接点可視化の主軸は Topic とし、Concept は厳格な語彙フィルタを伴う補助的手段として位置づけるのが妥当である。Social Sciences 領域の日本語文献は、出版社分布（国内学会中心）からみて機械翻訳メタデータではなく日本語原著の比重が相対的に高いことが示唆される。ただし、OpenAlex における日本語の人文・社会科学文献の収録は限定的で、伝統的人文学は数十件規模にとどまる。書籍の割合は当該分野の方が全体より相対的に高いものの、Hicks（2004）が指摘する同分野における書籍の重要性を踏まえると、なお十分とは言い難い。

この結果は、OpenAlex における日本語文献全体の質的担保と多様性の複雑な状況を示している。MAG 除外により、一般雑誌記事の混入という質的問題は解消されたものの、機械翻訳メタデータと真の日本語原著論文の識別という課題が残される。また、書籍や地域出版物の収録が極めて限定的であることから、特に人文学・社会科学分野における多様性の確保について、分野別（文学・歴史・地域研究等）の媒体構成に沿った収録設計と、書籍・地域誌の積極的な包含が求められる。

参考文献

Eisenstein, E. L. (1980). The Printing Press as an Agent of Change. *The Printing Press as an Agent of Change*. https://doi.org/10.1017/CBO9781107049963

Stracke, C. M. (2020). Open Science and Radical Solutions for Diversity, Equity and Quality in Research: A Literature Review of Different Research Schools, Philosophies and Frameworks and Their Potential Impact on Science and Education. *Lecture Notes in Educational Technology*, 17–37. https://doi.org/10.1007/978-981-15-4276-3_2

UNESCO. (2021). *UNESCO Recommendation on Open Science*. https://doi.org/10.54677/MNMH8546

Hicks, D. (2004). The four literatures of social science. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems* (pp. 473–496). Springer. https://doi.org/10.1007/1-4020-2755-9_22

Facing the Future: European Research Infrastructures for the Humanities and Social Sciences - ALLEA. (2014). Retrieved March 28, 2025, from https://allea.org/portfolio-item/facing-the-future-european-researchinfrastructures-for-the-humanities-and-social-sciences/

沼尻 保奈美, Ismael Rafols, André Brasil, 林 隆之, 林 和弘（2025）.「バルセロナ宣言とは何か―研究情報のオープン化に向けて―」.STI Horizon. NISTEP. https://doi.org/10.15108/stih.00403

謝辞

本研究は、文部科学省「人文学・社会科学のＤＸ化に向けた研究開発推進事業（人文学・社会科学研究におけるデータ分析による成果の可視化に向けた研究開発）」の支援を受けて実施した。