研究データの検索ツール

池内　有為*

キーワード：データ検索，サーチエンジン，オープンサイエンス

1. はじめに

あっという間に 2 年目の本連載。執筆陣に尾城孝一さん（国立情報学研究所）を迎え，佐藤翔さん（同志社大学）の筆致は相変わらず冴え渡り，大船に乗った気持ちである——というのはまあ虚勢で，常々「このネタでいいのかしら」と思いながら書いている。勇退された林豊さん（現・国立情報学研究所）をはじめ，皆さまからの忌憚のないご意見をお待ちしております。

さて，研究データの公開が拡がり分野横断で検索できるツールも増えてきた。そこで今回は主要な 4 つのデータベース，すなわち Clarivate Analytics 社の Data Citation Index（DCI）[1]，DataCite[2]の検索（DataCite Search）[3] 注1)，Elsevier 社の DataSearch（ベータ版）[4]，そして Google Dataset Search（ベータ版）[5]を取り上げたい。2 章ではそれぞれの概要を，3 章ではキーワード検索の結果と使い勝手を，4 章ではタイトル検索の結果を紹介する。

2. 研究データ検索ツールの概要

まず，4 つのツールで検索できる対象や特徴について概観する。結論から言うと，2019 年春の現時点では 4 ツールの収録範囲が結構異なるので，それぞれの特徴を知っておくと良いと思う。

2.1 Data Citation Index（DCI）

2012 年に Thomson Reuter 社（現 Clarivate Analytics 社）が公開したデータベースで，Web of Science の一部として提供されている（有料）。その名が示す通りデータの引用索引であり，Science Citation Index（SCI）などと同様に，データとそのデータを引用した論文を検索することが可能である。

対象となるデータの種類は，データセット，データ研究（data study），ソフトウェアであり，随時リポジトリ単位でデータが追加されている（リポジトリ自体も検索対象である）。2019 年 3 月末現在，推定 888 万レコードが収録されている。

＊いけうち　うい　文教大学文学部英米語英米文学科
〒343-8511　埼玉県越谷市南荻島 3337
E-mail: ikeuchi@koshigaya.bunkyo.ac.jp
orcid.org/0000-0002-5680-1881
（原稿受領　2019.5.8）

2.2 DataCite Search

研究データを対象としたデジタルオブジェクト識別子（Digital Object Identifier, DOI）の登録機関（Registration Agency, RA），つまりは元締めである DataCite による検索サービス。DataCite Metadata Search（ベータ版）を経て提供が開始された。

DataCite の統計によれば，2019 年 4 月末現在，DOI を発行する 130 機関によるメタデータのうち発見可能（findable）なもの，すなわち DataCite Search で検索できると推測されるのは 1,454 万レコードである [6]。

2.3 Elsevier DataSearch（ベータ版）

2016 年 8 月に Elsevier 社が公開した研究データを探すためのデータベース。その特徴は，ScienceDirect と arXiv（プレプリントサーバ）に掲載された論文の図表や補足資料を検索対象に含めている点である。このほか，同社が選択したデータリポジトリやリポジトリのメタデータを収集して索引付けを行っている。絞り込み条件で「データリポジトリ（Data Repositories）」を選択すればリポジトリのデータのみ，「論文リポジトリ（Article Repositories）」を選択すれば補足資料のみを表示することができる。

データ件数は不明だが，収録リポジトリはすべて列挙されている。FAQ（よくある質問）によれば，リポジトリの選択基準はユーザ数やインデックス作成の容易さとのこと。

2.4 Google Dataset Search（ベータ版）

2018 年 9 月に Google が公開したデータベース。対象は「研究」データに限定しておらず，csv や画像など，データセットとみなされるものを幅広く検索できる。ただし，Google Dataset Search が検出するのは schema.org の Dataset マークアップや W3C の Data Catalog Vocabulary（CDAT）に基づく構造化データをもつデータセットである。畢竟，対応していないデータ（リポジトリ）は，Google Dataset Search で検索されないことになる（！）。Google による 2017 年の構想発表から Figshare などによる迅速な対応，そして Dataset Search 公開までの経緯は，オープンサイエンス基盤研究センターのブログで船守氏が解説し

ている[7]。

3. キーワード検索

データを検索する状況＜その 1＞として，“研究や調査に使うためのデータを探してみる”場合を想定して，「library statistics」などをキーワードに無料の 3 ツールで検索した。ちなみに日本語のキーワードで検索すると DataCite はごくわずかしかヒットせず，Elsevier はエラー，Google でヒットするのはほとんどが中国語のデータであった（船守氏のブログが書かれた時点とあまり変わらず？）。

検索画面は 3 ツールともシンプルで差がなく，レスポンスはどれも速い（Web サイトの表示速度のスコアは微妙に差があるが，体感できるほどではない）。ここでは「machine △learning△dataset」で検索した結果を示す。

DataCite は 1,999,773 件ヒット。登録年，データ形式，データセンター（リポジトリ）で絞り込みが可能。一覧を見ていくと，クリエイティブ・コモンズマークが表示されるデータもある。データ引用を推進する組織だけあって，コピペ用の引用情報を APA，Harvard，BibTeX など 8 種類の形式で表示できる。

Elsevier は 51,978 件ヒット。うちデータリポジトリのデータは 29,818 件，論文リポジトリのデータは 22,160 件。このほかデータ形式，リポジトリ，日付で絞り込みが可能。一覧表示画面では，図表などのデータやメタデータやのプレビューが表示できる。

Google は「100 件以上の検索結果が見つかりました」（＝ヒット件数不明）。ランキングが絶妙で[注 2]，まず Kaggle Datasets（データサイエンスや機械学習のプラットフォーム），次いで Kaggle の面白そうなデータ（Pokémon for Data Mining and Machine Learning など）を 4 件，そして CERN の ATLAS ヒッグス粒子機械学習チャレンジからのデータセットと続く。ただし絞り込みや並べ替えはできないため，適宜キーワードをみつけて追加していくしかない。唯一スマホに対応している。

4. タイトル検索

データを検索する状況＜その 2＞として，“引用文献などで知ったデータを探す”場合を想定した。具体的には DCI でよく引用されているデータを調べて，そのデータが無料の 3 ツールで検索できるどうかを試すことにした。

まず，DCI で 2000 年から 2019 年までのデータセット，データ研究，ソフトウェアを検索して，それぞれ被引用数が多い順にソートした。上位 5 件を選択しようと考えたが，上位のデータは同じリポジトリに収録されていることが多いため，当該リポジトリを収録していない検索ツールはノーヒットばかりということになってしまう。そこで，1 リポジトリにつき 1 件という条件でデータを選択した[注 3]。

表 1 に検索結果を示す。見出しは左から，DCI の順位，DOI の有無，各データベースのヒット状況，DCI の被引用数，Google Dataset Search の被引用数を示している。検索結果の 10 位以内に検索対象のデータが表示された場合はヒットしたとみなして，TI（タイトル検索）または TI-M（タイトル完全一致検索）と記した。なお，DCI のレコードとは別のリポジトリに登録されたデータであっても，内容が同じならばヒットとみなした。

表 1　タイトル検索の結果

順位	DOI	DataCite	Elsevier	Google	Ci	G-Ci
dataset						
1	有*	TI-M	TI	TI	325	0
53				TI	121	0
91				TI-M	64	0
122					50	
134					47	
data study						
1				TI	643	298
4					268	
8		TI-M	TI	TI	206	6
13			TI		185	
52	有	TI-M	TI	TI	150	14
software						
1					527	
23	有	TI-M	TI		26	
99					6	
118	有	TI-M		TI	5	15
118	有	TI-M		TI	5	0
合計		6	5	8		

Ci=Data Citation Index被引用数
G-Ci=Google Dataset Search被引用数
TI=タイトル検索
TI-M=タイトル完全一致検索
*DOI変更

サンプル数は少ないが，改めてデータ引用の重要性を認識した。DOI が記されていれば容易に元データにアクセスできるが，ない場合は高被引用データであってもデータ検索ツールでみつけるのは（今のところ）難しそうである。また，収録データはツールによって異なり，被引用文献数は DCI と Google で乖離がみられた。教科書的には“広範にデータを探す場合は，複数のツールを併用する必要がある”と言わざるを得ないだろう。

5. おわりに

本稿で紹介したデータ検索ツールは随時データや機能を追加しており，いずれもフィードバックを受け付けている。データの発見と活用のために最適な検索ツールのあり方を，データの公開者や利用者を巻き込みながら，まさに「いま」検討している真っ最中という印象を受けた。

さて，本連載は 2019 年度から即時オープンアクセスとなった。エンバーゴ付きだった昨年度は，筑波大学のつく

ばリポジトリに登録していただいていた。刊行後すぐに反響を得ることができたのは，ひとえにリポジトリ担当の皆さまのおかげであり，オープンサイエンスによる学術情報の循環は，こうした現場の皆さまに支えられているのだと実感した。この場を借りて，心よりお礼を申し上げたい。

註・参考文献

注 1) 検索画面に名称はついていないが，本稿では機関名と区別するためにページタイトルである「DataCite Search」と記す。
注 2) パーソナライズ検索が行われているかどうかは不明。別のネットワーク，別の地域にいる人にも検索してもらったが，似たようなランキングとなった。なお，検索当時は映画「名探偵ピカチュウ」の封切直後であった。
注 3) データ研究のうち，DCI の書誌が不完全であった 1 件は除外して，次点のデータを検索対象とした。

1) Data Citation Index. https://clarivate.com/products/web-of-science/web-science-form/data-citation-index/, (accessed 2019-05-06).
2) 福山樹里．DataCite：国立図書館×DOI×研究データ．カレントアウェアネス，2015，no.324，p.8-11. http://doi.org/10.11501/9396324, (参照 2019-05-06).
3) DataCite (Search). https://search.datacite.org, (accessed 2019-05-06).
4) Elsevier DataSearch beta. https://datasearch.elsevier.com/, (accessed 2019-05-06).
5) Google Dataset Search beta. https://toolbox.google.com/datasetsearch, (accessed 2019-05-06).
6) DataCite Statistics. https://stats.datacite.org, (accessed 2019-05-06).
7) 船守美穂．“グーグル，オープンデータのための検索エンジンを発表”．RCOS 日記－miho チャネル．2018-09-08. https://rcos.nii.ac.jp/miho/2018/09/20180908/, (参照 2019-05-06).

Series: Current trend of open science: Review of Research Data Search Tools.

Ui IKEUCHI (Bunkyo University, Faculty of Language and Literature, Department of English Language and Literature, 3337 Minami-Ogishima, Koshigaya, Saitama 343-8511)

Keywords: Dataset Search / Search Engine / Open Science