映像情報メディア学会誌
第 103 号　特集「私の研究開発ツール」

文献検索ツール

山里 敬也、西野 文人、加藤 直人、千村 保文

アブストラクト　主な学術文献データベースを紹介し、その内容を分析する上で重要な研究者 ID やメタデータなどの考え方を解説する。また、メタデータに基づくデータベース例を紹介するとともに、文献管理ツールの事例と課題について述べる。

キーワード　学術文献データ、研究者 ID、メタデータ、Linked Data、LOD

1.はじめに

研究開発において先行研究例の調査のために学術文献の調査は欠かせない。そこで、世界中の学会や企業において、学術文献に関するデータベースが整備されている。しかし、その目的や登録されているデータや登録形式は様々であり、文献検索の目的によって使い分ける必要がある。

本稿では、主な学術文献データベースを紹介し、学術文献データの検索を取り巻く環境や Linked Data、LOD などのキーワードと文献管理ツールについて解説する。

2.学術文献検索システム

表 1 に主な学術文献検索システムを示す。

表１　主な学術文献検索システム

システム名称	分野	システム提供者	システムの概要
Arnetminer	計算機科学	清華大学（中国）	Arnetminerは、中国国家ハイテクR&Dプログラムと中国科学技術財団が資金を提供した社会的影響分析、ソーシャルネットワークランキング、ソーシャルネットワーク抽出の研究プロジェクトとしてスタート。ソーシャルネットワーク分析により、研究者、ジャーナル、会議予稿の関連性を導くデータマイニングができる。これにより、エキスパート検索、地理的検索、査読者の推奨、関連検索、コース検索、学術業績評価、トピックモデリングなどのサービスを提供している。
arXiv	物理学、数学、計算機科学、量的生物学、計量ファイナンス、統計学	コーネル大学（米国）	1991年にスタートしたプレプリントサーバ。数学と物理学の多くの分野では、ほぼすべての論文がarXivに自己アーカイブされている。登録文献数は2014年末には100万を突破。月に10万件以上の登録がある。
ACM Digital Library	計算機科学，工学	ACM（Association for Computing Machinery）	計算機科学分野で最も権威のあるACMが提供するデジタルライブラリー。IEEEがエレクトロニクスや通信分野の工学に強いのに対し、数学的な理論計算機科学のような分野もカバーする。
BASE: Bielefeld Academic Search Engine	全分野	ビーレフェルト大学（ドイツ）	オープン・アーカイブ・イニシアティブ・プロトコル（OAI-PMH）を実装する機関リポジトリおよび他の学術的デジタル・ライブラリからOAIメタデータを収穫し、検索のためにデータを正規化および索引付けしている。

CiNii	全分野	国立情報学研究所（NII）	NII が運営する学術論文や図書・雑誌などの学術情報データベース。2017 年 3 月に電子図書館（NII-ELS）事業の終了に伴い、有料コンテンツの提供（個人 ID、従量課金、機関定額制）も終わり、現在はオープンアクセスの文献のみが閲覧できる。
CiteSeer	計算機科学	ペンシルベニア大学（米国）	CiteSeerx（旧 CiteSeer）は、計算機および情報科学分野の文献検索エンジンで Google Scholar や Microsoft Academic Search などの学術検索ツールの前身として知られている。
The Collection of Computer Science Bibliographies	計算機科学	Alf-Christian Achilles	1993 年にスタート。計算機科学の文献データベースとしては恐らく最も古い。
Dimensions	全分野	出版社の合同ポートフォリオ（ReadCube, Altmetric, Figshare, Symplectic, DS Consultancy, ÜberResearch）	Digital Science 社によって開発された linked research knowledge システム。世界中の 100 を超える主要な研究機関と協力し、ユーザーが最も関連性の高い情報を見つけてアクセスするためのプラットフォームを提供する。
dblp computer science bibliography	計算機科学	トリーア大学（ドイツ）	1993 年にドイツ・トリアー大学で始まる。計算機科学に関する雑誌論文、会議論文、その他の出版物を 366 万件以上保持し、計算機科学に関する主要なジャーナルおよび国際会議予稿集をカバー。

Google Scholar	全分野	グーグル	Google が提供する文献検索サービスで、ネット上に散らばっている同一論文をまとめて表示する。
IEEE Explorer	計算機科学、電気工学、電子工学	IEEE	IEEE Xplore は、IEEE が発行する論文誌および国際会議予稿に加え、他のパートナーが発行する文献を含む検索システムであり、450 万件を超えるドキュメントを保持しており、毎月約 20,000 の新しい文書が追加されている。計算機科学、電気工学、電子工学、および関連する分野では世界で最も引用されている出版物を持つ。
IngentaConnect	全分野	Ingenta	13,500 件以上のジャーナルデータベースと、ジャーナルと電子ブックを含む 450 万件の記事を掲載。
J-Stage	全分野	独立行政法人科学技術振興機構(JST)	JST が運営する電子ジャーナルの無料公開システムで、2018 年 10 月現在、2,693 ジャーナルの 4,688,103 論文を収録している。
Genamics JournalSeek	全分野	JournalSeek	5400 を超えるジャーナルのオンラインデータベース。
JSTOR: Journal Storage	全分野	JSTOR	1995 年に設立されたデジタルライブラリー。約 2,000 ジャーナルのフルテキスト検索を提供。
Microsoft Academic	全分野	マイクロソフト	マイクロソフト・リサーチによって開発された文献検索エンジン。セマンティック検索をサポート. 3 億 7,500 万のエンティティがあり, そのうち 1 億 7,000 万は学術論文である.
国立国会図書館オンライン	全分野	国立国会図書館	国立国会図書館検索・申込オンラインサービス(略称:国立国会図書館オンライン)は、国立国会図書館の所蔵資料及び国立国会図書館で利用可能なデジタルコンテンツを検索し、各種の申込みができるサービ

			ス。2017 年 12 月に終了した NDL-OPAC(National Diet Library Online Public Access Catalog)の後継．書誌情報をダウンロードする機能に特化した「国立国会図書館書誌提供サービス(略称:NDL-Bib)」とは別．
OAIster	全分野	OCLC (Online Computer Library Center)	2002 年にミシガン大にてスタートした検索サイト．OA ジャーナル数は世界最大．
Science.gov	全分野	米国エネルギー省科学技術情報室	米国政府の科学技術情報の検索システム。現在、14 の連邦科学機関からの 38 以上のデータベースから 2 億ページの科学情報を検索できる。
ScienceOpen	全分野	ScienceOpen	arXiv，PubMed，SciELO の文献探索ができ，ORCID もサポートしている．
Scopus	全分野	エルゼビア	5,000 以上の出版社による約 22,000 文献を検索できる．検索機能対象には特許データベースも含み，著者の略歴・著作数・書誌データ・引用数や引用先も取れる．著者別 ID はオープンソースの電子識別子 ORCID と統合できる。
Semantic Scholar	計算機科学，生物医学	Allen Institute for Artificial Intelligence	マイクロソフト社共同創業者のポール・アレンが設立した人工知能研究所で開発された検索システム。機械学習、自然言語処理、マシンビジョンを使用して、従来の引用分析方法に意味分析レイヤーを追加している。Google Scholar と PubMed と比較して、最も重要な論文をすばやく

			ハイライトし、それらの間のつながりを特定するように設計。
Sparrho	全分野	Sparrho	AI による文献検索支援システム。関連する文献を集約し推薦する機能を持ち、ユーザーの入力支援による機械学習や文脈分析を通じて、適切なコンテンツや関連する科学分野にわたる偶然の発見を可能にしている。
SpringerLink	全分野	シュプリンガー	シュプリンガーが提供する文献データベース。
Web of Science	全分野	クラリベイト・アナリティクス	約 12,000 文献を検索できる。インパクトファクターの計算根拠でもある。
WorldCat	全分野	OCLC (Online Computer Library Center)	OCLC に参加する 71,000 以上の図書館の蔵書を目録化した総合目録

この表は文献[1]のリストの内、映像情報メディア学会と関連の深い、計算機科学、工学、電気工学、電子工学、また全学術分野（Multidisciplinary）の文献検索システムに、J-Stage を加えたものである。これらのシステムは学術論文出版社が提供するもの、大学・研究機関が提供するもの、民間企業等が提供するものに大別できる。また、文献のタイトル、著者、要約などの書誌情報については無償で検索できるが、本文閲覧については定期購読会員に限定されているものが多い。オープンアクセスジャーナルを対象にした文献検索システムもあり、これらのシステムでは本文閲覧も無償でできる。

これらの文献検索システムに加え、プレプリントサーバも文献検索システムとして利用されることが多い。ここで、プレプリントサーバとは、査読のある学術雑誌の投稿と同時に、あるいは掲載が決定した著者がその原稿をアップロードできるサーバを指し、いち早く文献を公表できる。また、ジャーナル掲載論文（ダウンロードは有償）と同じ内容の著者最終原稿もあり，これを無償でダウンロードできることも魅力である。同様の仕組みと

しては、大学が提供している機関レポジトリもある。これらのシステムが提供する文献は無償で閲覧できるため、オープンアクセスジャーナル同様に扱われることが多い。余談だが、arXiv はロシア人数学者グリゴリー・ペレルマンがポアンカレ予想を解決した論文を投稿したことでも注目された。ペレルマンはこの功績により 2006 年のフィールズ賞を受賞したが，本人は辞退している。

最近のトレンドとしては、AI による文献検索支援を行うシステムがある。たとえば、Arnetminer はソーシャルネットワーク分析により、研究者、ジャーナル、会議予稿の関連性を導くデータマイニングができ、Microsoft Academic はセマンティック検索をサポートしている。Semantic Scholar では、機械学習、自然言語処理を使用して、従来の引用分析方法に意味分析レイヤーを追加している。Sparrho は関連する文献を集約し推薦する機能を持ち、ユーザーの入力支援による機械学習や文脈分析を通じて、適切なコンテンツや関連する科学分野にわたる偶然の発見を可能にしている。

3. モノ・コトを扱うツールと技術

読みたい論文のタイトルがわかっていてその論文自身を探したい場合や、あるキーワードから関係のありそうな論文を適当に数編探したいようなときは Google のようなキーワードによる文書検索で事足りる。それでは、学術文献の検索にも Google があればそれで十分であろうか？上記のような目的だけであるならば、Yes と言えるかもしれないが、研究開発の様々なケースを考えると論文本文中の文字列を対象とした文書検索では不十分である。例えば、ある著者の論文を多く拾いたいというような特定の状況を設定した検索であるとか、トレンドを分析したり特定の分野の専門家を探したいというようなときには、文字列で検索・分析するのではなく、論文のメタデータのモノ・コトで検索・分析できるようになっていることが必要である。メタデータとは、データそのものではなく、そのデータを表す属性や関連する情報を記述したデータのことであり、論文のメタデータとしては、論文のタイトルや、要約、著者名、著者の所属、掲載雑誌名、発行年などの情報がある。論文の著者や発行年などは、文字列として検索・処理するのではなく、著者という人物としてあるいは発行された年として検索・処理できることが望まれる。そこで、このような論文のメタデータ、すなわち、著者や（著者の所属）機関、発行年などは、文字列で管理するのではなく、それぞれの属性が持つタイプに従って、簡単に取り扱えるようになっていれば、そのデータから必要な情報を集計して、分析したり、好きな形に変形することができるので便利である。この節では、論文の様々な情報をモノ・コトとして扱っているツールやそれを支える技術について述べる。

3.1.　文献・研究者の検索・一覧サービス

ここでは文献や研究者を探索したり、そのメタデータを閲覧できる代表的なサービスを紹介する。

3.1.1.　研究者総覧 researchmap

researchmap は科学技術振興機構(JST)が運営する研究者データベースであり、26 万人以上の研究者情報を有している。researchmap では研究者は基本的情報に加えてプロフィールや経歴、業績情報などを登録・表示できる。また Amazon, Arxiv, CiNii, dblp, J-GLOBAL, ORCID などの外部データベースから researchmap 業績項目へ情報を取り込むこともできる。さらに、研究者情報を Web API を介して自機関データベースに取り込んだり、自機関データベースから researchmap へリンクを張って利用するなどが行われている。なお、2019 年度にサービス開始が予定されている次期 researchmap では、標準化と名寄せの強化によって、機械可読性の向上が予定されているとのことである。

3.1.2. J-GLOBAL(http://jglobal.jst.go.jp)

J_GLOBAL は JST 内外の科学技術情報（国内外の文献約 4650 万件、特許約 1300 万件、機関約 53 万機関、研究者約 27 万人、研究課題約 6 万件、科学技術用語約 33 万語、化学物質約 372 万件など）を相互に関連付けたサービスである。J-GLOBAL では目的別検索として、研究者、文献、化学物質などを所属や発行年、各種 ID などの項目を指定して検索できるようになっている。

3.1.3. DBLP

DBLP(https://dblp.org/)は、コンピュータ科学に関する書誌メタデータを提供するサイトであり、2018 年 10 月現在 430 万件を超える出版物を有している。現在 DBLP には後述の ORCID 情報が定期的に追加されるようになっていて、ORCID を持つ著者の割合は約 10.9%までに上昇しているとのことである。DBLP では論文メタデータの検索サイトが用意されており、例えばキーワードを指定して検索すると関連する論文がリストアップされる。また、著者や年、論文タイプなどで絞り込むことも可能である。さらに著者などはリンクになっており、ある著者の論文のリストを求めることもできる。

3.2. 研究者 ID

研究者と業績を結びつけようとしても、同姓同名の研究者の存在や名前の表記の不統一（異体字や外国語表記）、略記（アルファベットでイニシャルだけの記述）、誤り（つづり間違い漢字間違い、筆頭著者以外ではときたま見かける）などにより、この作業はとても大変なものであった。研究者を名前で管理するのではなく、同一人物には一つの ID を与え、別の人物には異なる ID を与えることで、著者等の名前のあいまい性が解消し、研究者と業績が結びつけやすくなる。文献に対する永続的な識別子としては、DOI(Digital Object Identifier)が利用されているが、一方研究者に対する識別子としては、これまではトムソン・ロイター社の Researcher ID やエルゼビア社 Scopus の Author ID、国立情報学研究所

の CiNii の ID、そして各学会の会員番号などがあり、これらは別々に管理されているので、これでは横断的に研究者と業績を結びつけることはできなかった。そんな中、世界中の研究者に一意の識別子を付与することで、これによって研究者の負担を軽減する学術情報基盤を構築することを目的としたのが ORCID(Open Researcher and Contributor ID)である。

ORCID は、非営利団体 ORCID Inc.が管理する名前や所属が変わっても永続的に管理・利用できる識別子である。ORCID は 2012 年 10 月にサービスを開始しており、2017 年 7 月上旬に ORCID を利用している研究者は 500 万人を超え、2018 年 10 月現在は約 540 万人（このうち日本からの登録者は約 8 万人）の研究者が ORCID に登録しており(https://orcid.org/statistics)、世界 44 カ国の研究機関、出版社、研究助成団体などの約 900 の機関が ORCID メンバとして加入している。因みに、研究者は ORCID アカウントを無料で取得し、すべての情報をコントロールできるので、まだ ORCID アカウントを取得していない方は、すぐに取得すると良いでしょう。現在、研究者自身によって入力された情報に加え、ORCID メンバー機関が追加する情報が蓄積・流通している。さらに論文投稿時に ORCID の入力を義務化するジャーナルも増加しており、今後は新しい論文が出るたびに著者の実績として自動的に追加されたり、引用された論文の著者への通知サービスなどが運用されたりすることが期待される。今後、論文や著者、研究機関の ID 化によって、研究活動は ID のネットワークとして表現されるようになることが期待されている。

3.3 知識インフラ

ここでは、論文や研究者等の情報を知識として活用するための知識インフラについて述べる。

3.3.1. Linked Data とは

近年有用な様々なデータを Web に皆で公開して共有しようという動きが進みつつある。Linked Data は、Web でデータを再利用しやすいように、共有・公開するための技術や方法論である。Linked Data では、すべてのモノ・コトに ID を付与することで、文字列ではなく、この ID で管理する。具体的には ID として URI(Uniform Resource Identifier)あるいはその国際化版である IRI(Internationalized Resource Identifier)が使用される（本稿では以下 IRI で表記することにする)。そして、そのモノ・コトの内容を確認できるように http プロトコルを使い（すなわち通常のブラウザでアクセスできる)、IRI を参照したときにはそのモノ・コトの情報を獲得できるようにする。さらに多くの情報を発見できるように他の情報への IRI リンクを含めている。情報の記述の仕方は規定されてはいないが、通常は標準の技術である RDF(Resource Description Framework)を使って記述する。Linked Data の検索や分析にはこれも標準である SPARQL と呼ばれるクエリー言語を使うことができ、この SPARQL クエリー言語を使ってそのデータを検索することを可能とするサイト（SPARQL エンドポイントと呼ばれる）を公開しているものもある。

主な Linked Data とその SPARQL エンドポイントとしては、dbpedia（wikipedia のインフォボックスと呼ばれる部分などを Linked Data 化したもの）のグローバル版(http://dbpedia.org/sparql)と日本語版(http://ja.dbpedia.org/sparql)や、wikidata（データ共有サイト。wikipedia のデータ版にあたる）（https://query.wikidata.org/）、NDL Authorities（国立国会図書館典拠データ（http://id.ndl.go.jp/auth/ndla/sparql）、e-Stat（日本政府統計データ）（http://data.e-stat.go.jp/lod/sparql/）などがあり、様々な情報が相互にリンクしあっている。例えばヘルスケア分野を見てみると、論文、薬剤、副作用事例、病気、臨床試験、タンパク質、遺伝子、医療保険、制度などのデータが、かつては書式や名称表記も異なっていて相互にリンクも無かった、現在はそれぞれが Linked Data 化され、お互いの情報がリンクで容易に参照可能になっている。

Linked Data のメリットとしては、データのフォーマットや意味解釈が標準化されているので、ツールの共通化を図りやすいこと、Web 上での情報公開・共有の仕組みがあり、それぞれの機関が独立に分散して Linked Data を立ち上げることができること、厳密な RDF に基づいてデータを作成することで機械解釈可能であり分析・推論が可能なことなどがあげられる。Linked Data の作成には初期投資が必要だが、論文のメタデータスキーマを整備し、メタデータを作成しておくことは、論文のメタデータをデータとして活用するためには、必要な投資であろう。

3.3.2. J-GLOBAL Knowledge

JST は第 4 期科学技術基本計画（2011 年）の知識インフラ構想を受け、「JST 知識インフラ」として、J-GLOBAL のデータ提供口の一つとして J-GLOBAL Knowledge を Linked Data としての構築を進めていた。その中で化学物質データ（日化辞 RDF データ）がクリエイティブ・コモンズ・ライセンス CC BY で公開されている。しかし残念ながら化学物質データの J-GLOBAL Knowledge の中のその他のデータ（文献や特許、研究者などの情報）は未だ公開されていない。

3.3.3 DBLP EXPLORER

DBLP では最新データを xml で提供しているので、これを使って Linked Data を構築した試みがある。dblp.rkbexplorer.com は RKB Explorer イニシアティブが運営するサイトであり、DBLP の SPARQL エンドポイント(http://dblp.rkbexplorer.com/sparql/)を提供している。しかし、残念ながらその中身のデータは 2013 年 2 月に dblp からデータを取り込まれたのを最後にその後データは取り込まれていない。

3.3.4 LOD4ALL

LOD4ALL（http://lod4all.net/）は富士通研究所が運営するサイトであり、世界中で公開されている Linked Open Data (LOD)を収集して一括検索することを可能にする LOD の活

用基盤であり、2600 以上のデータセットに対する SPARQL インタフェースを提供している。またこの上に構築されている LOD4ALL Frontend（http://lod4all.net/frontend/）は、LOD の各エンティティを様々な観点で可視化してナビゲーション・閲覧を可能とするサービスであり、前述の日化辞情報や、DBpedia をベースとした生物辞典、国税庁の法人番号や経産省の法人インフォをベースとした企業情報分析 Web アプリなどを公開している。

4. 文献管理ツール

文献検索は、新たな文献を収集するためにだけではない。それまでに自身が収集してきた文献を検索する場合がある。収集した文献が大量になってくると、必要な文献を探し出すことは容易ではない。そこで、検索しやしいように収集した文献を管理する必要に迫られる。

そのような文献管理ツールとして、

Endnote（http://www2.usaco.co.jp/）、

Evernote（https://evernote.com/intl/jp）、

Mendeley（https://www.mendeley.com/）などがある。

それぞれ、有償／無償、使い勝手など利点や欠点があるが、ここでは無償（個人では 2GB まで）で比較的使いやすいと感じている Mendeley を紹介する。ただし、Mendeley に関してはすでに優れた解説[2]があるので、ここでは概要にとどめる。

Mendeley は Elsevier 社の文献管理ツールである。デスクトップ版と Web 版があるが、デスクトップ版と Web 版は同期させることができるので、デスクトップ版を利用するほうが後述するように便利である。デスクトップ版は Windows、Linux、Mac に対応している。

Mendeley Desktop（デスクトップ版）を立ち上げると、「All Documents」の画面に入り、登録済みの文献が表示される。新たに文献を登録するのはいたって簡単で、文献（PDF ファイル）を画面上にドロップすればよい。すると、Mendeley は書誌情報（文献タイトル、著者、発行誌、・・・）を自動的に登録する。しかし、日本語の文献は自動登録できないので、書誌情報を人手で入力する必要があるという欠点がある。また、文献（PDF ファイル）も同時に登録されるので、個人が無償で利用する場合には 2GB の制限にかかりやすい。ただし、文献をデスクトップ上に保存するように設定変更すればこの問題は回避できる。

5. おわりに

主な学術文献データを紹介し、文献データを効果的に分析するためにメタデータや Linked Data の考え方を紹介した。また、文献管理の上での課題とツールの活用について述べた。情報通信分野の研究は多くの分野への応用領域が広がっており、分野横断での検索が必要である。しかし、本稿で紹介したとおり、学術文献データの多くは分野ごとに構築されており、かつメタデータ化されている例はまだ限られる。効率的な研究開発を促進する上で、分野横断的なメタデータ化された Linked Data による学術文献データの構築が

重要と考える。

参考文献

[1] List of academic databases and search engines, Wikipedia, https://en.m.wikipedia.org/wiki/List_of_academic_databases_and_search_engines (2018 年 10 月 5 日)
[2]高取 憲一「論文情報管理･共有ツール“ Mendeley ”」映像情報メディア学会誌, Vol.70, No.3, p.320-323, 2016.

山里敬也
1993 年慶大大学院博士課程了．博士（工学）．現在，名大教養教育院教授．2006 年 IEEE Communications Society 2006 Best Tutorial Paper Award, 2014 年電子情報通信学会会長特別表彰を受賞．可視光通信，ITS，e ラーニングなどの研究に従事．電子情報通信学会，IEEE および本会正会員.

西野 文人
1981 年，東京工業大学大学院修士課程修了．同年，㈱富士通研究所入所．以来、機械翻訳、情報検索、ナレッジグラフなどの自然言語処理・知識処理に関わる研究開発に従事。現在、同研究所人工知能研究所特任研究員．ORCID: 0000-0001-7368-4923

加藤 直人

1988 年，早稲田大学大学院修士課程修了．同年，ＮＨＫ入局，放送技術研究所に勤務．この間，ＡＴＲ音声翻訳通信研，ＡＴＲ音声言語コミュニケーション研に出向．現在，スマートプロダクション研究部上級研究員．博士（情報科学）．機械翻訳，自動要約などの研究に従事．正会員

千村　保文

1981 年,日本大学理工学部卒業.

同年，沖電気工業株式会社入社，VoIP システムの研究開発，標準化活動に従事．現在，経営基盤本部にて政策動向調査，イノベーション教育を担当．電子情報通信学会および本会正会員.