多摩地域の図書館の蔵書データを斬新な方法で有効活用します

ビッグデータで見えてくる多摩地域図書館

吉本龍司（よしもと りゅうじ）
※株式会社カーリル 代表取締役・エンジニア

すべての図書館をつなぐ

図書館蔵書検索サイト「カーリル」は2010年3月にサービスを開始した。全国の図書館で分散して管理されている所蔵情報をウェブスクレイピング技術により効率的に統合し、使いやすいウェブサービスを提供することに成功したのである。現在、全国6，600以上の図書館・図書室の検索に対応しており、広告収入により運営されているため無料で利用することができる。公共図書館のカバー率は93パーセントを超えた。

このサービスの開発にあたっては「すべての図書館をつなぐ」ことを目指してきた。特に、ウェブ検索がすべての〝調べる〟ことの起点となっている今、ウェブから図書館への入り口として機能するサービスを作りたかった。最近は、行政のデータを活用して新たなビジネスを創出するというオープンデータの活用事例としても注目されるようになっている（開発当初はそういった概念がまだ日本に浸透しておらず当然意識もしていなかったが、現在では政府が成長戦略の一環として位置づけている）。

一方で、忘れられない事件もある。折しもカーリルのサービス立ち上げと同時期に発生した岡崎市立中央図書館事件（通称、Librahack事件）である。カーリルと同じようにウェブスクレイピングにより図書館を便利に使おうとした利用者が、偽計業務妨害により逮捕された。のちに原因となったアクセス障害は、図書館システムの脆弱性であったことが判明した。

図書館APIと市民参加

この当時、全国の図書館が業務システムの「おまけ」として提供していたWe

bOPAC（蔵書検索システム）は多くの問題を抱えていた。使い勝手はもちろんのこと、機械的な大量アクセスに弱く、日常的にシステムダウンが絶えなかった。そういった状況の中で起きたのがLibrahack事件だった。俯瞰的に見れば図書館の常識と、ウェブの常識が技術レベルで乖離していたということだろう。

プログラムで図書館をもっと便利に使いたい、そう思う人は当然私たちだけではない。こうした声に応えるため、カーリルでは図書館APIを提供することにした。APIとは機械的にデータを提供する仕組みのことで、全国の図書館の所蔵情報に統一的なプロトコル（通信方式）でアクセスすることができる。図書館APIを活用して図書館の利用や読書活動を支援するアプリやウェブサービスは、500を超えた。スマートフォンアプリは特に競争が活発で、多くの図書館検索アプリが提供されている。図書館API を自らのサービスに活用する〝図書館〟も増えた。

多摩デポとの出会い

2014年10月からNPO法人共同保存図書館・多摩（多摩デポ）とカーリルとの共同研究が始まった。この取り組みは、カーリルの図書館APIを活用することにより、「ラスト・ワン・ツー」のデータベースを構築しようというものだ。多摩デポでは従来より、多摩地域の図書館がもつ図書の多様性を保つため、最後の1冊および2冊となる図書の保存を各図書館に呼びかけてきた。もちろん各図書館の書庫の収容率からみればこれも長期的には難しいため、各図書館が保存しきれない希少な資料を中心に、必要な資料を共同で保存し、いつでも提供できる仕組みとして「共同保存図書館」の設置に向けた運動を行っている。

一方で、多摩デポでは除籍（廃棄）する資料のラスト・ワン・ツー調査を代行するサービスも図書館向けに実施してきた。これは、都立図書館が提供する統合検索などを活用して、1冊ずつ調査するという大変な作業である。こうした除籍時の地道な確認作業をもっと簡単にできないだろうか。それが多摩デポからの最初の相談だったと記憶している。

データベースの構築にあたっては、技術開発とデータ処理をカーリルが担当し、その設計について多摩デポのメンバーと協議を重ねた。

大規模なデータ収集

まず取り組んだのは、図書館に追加の負担をかけることなく、データを収集することだった。全国を見渡すと、幾つかの県で、県立図書館が中心となって、県内自治体のラスト・ワン調査が始まっている。これらは各図書館から個別にデータの提供を受ける方式が主流であるが、データの受け渡しに手間がかかり、場合によってはシステム改修の必要性もあった。これらの課題はカーリルのAPIにより、直接図書館がウェブに公開している所蔵情報を収集することで解決できる。

カーリルでは基本的な書誌同定にＩＳＢＮを採用しているため、まずはＩＳＢＮが付与された本を対象として扱った。基本原理は簡単で、すべての図書館の所蔵情報を収集することができれば、最後の１冊を判定できる。国立国会図書館（ＮＤＬ）と国立情報学研究所（ＮＩＩ）の書誌データを活用し、ＩＳＢＮを持つ和書のリストを作成した。重複しているＩＳＢＮを１つにまとめると、およそ１８０万タイトルとなった。

２０１５年の１月からおよそ３ヶ月間で、多摩地区の29自治体の図書館システムを対象に総当りでデータを取得した。各図書館のエラー状況なども確認しながら、最終的に１日約３万冊のデータを取得するという頻度となった。単純計算で延べ５２２０万回機械的にアクセスしたことなる。

集めたデータを見てみると、ＩＳＢＮが間違っているケース、出版年が古い書誌についてはＩＳＢＮが入っていないケースなどが含まれている。これらのノイズを排除し精度を上げるには人力が必要となるが、ゴールであるラスト・ワン・ツーを判定したいという目的のうちでは、これらのノイズがそのままだったとしても、多くは安全側に倒れる（つまり、ラスト・ワンと誤判定される）ことになるようだ。

今回の調査により、ＷｅｂＯＰＡＣの速度が低下したとの指摘が１つの図書館からあったものの、このアクセスが直接的な原因となるシステム障害の報告はなかった。すでに図書館システムが提供するＷｅｂＯＰＡＣも、通常のウェブサービスと同様の技術水準に達しているということであり、こうした実績は今後の図書館の本格的なデータ活用において大きな一歩といえる。

ビッグデータ技術の活用

最近の技術トレンドとして、ビッグデータという言葉を聞いたことはないだろうか。大量のデータを超低価格で超高速処理する事が可能になったというのである。これは、技術的な進歩というよりむしろ社会システムの変化である。これまで、大量のデータを処理するためには、大規模なコンピューターシステムが必要となり、その導入コストは莫大なものとなった。しかし、クラウドコンピューティングの発達により、例えば、高性能なコンピューター１０００台を１秒だけ借りるといったことが可能となった。大規模なデータ処理が一気に現実味を帯びたのだ。

今回の調査で収集したデータを解析するためにＧｏｏｇｌｅが提供するビッグデータ処理の基盤技術である「Ｇｏｏｇｌｅ Ｂｉｇ Ｑｕｅｒｙ」を採用した。これにより先ほど述べた５２２０万回のアクセスを始め、ほとんどのデータ処理は１秒程度で処理することができ、大幅なコストダウンに成功した。

多摩地域の現状を把握する

結果を見てみよう。ＮＤＬとＮＩＩのもつ書誌のうち、多摩地域の29館の図書館が所蔵しているタイトルは56％にあたる約１０２万冊ということがわかった。

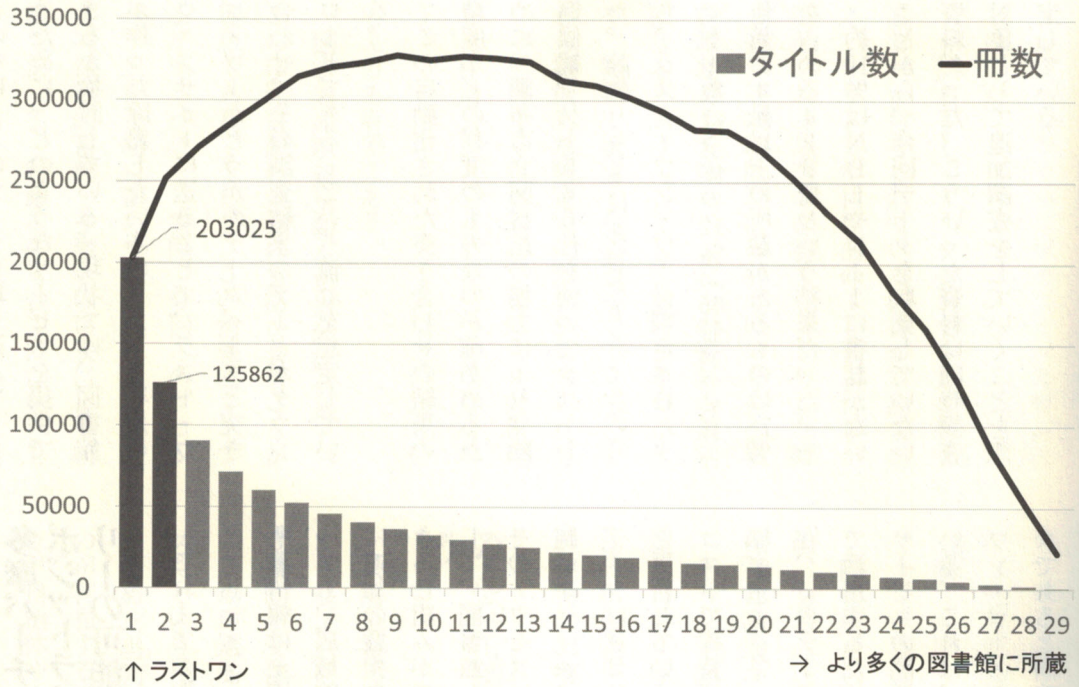

図1　所蔵冊数ごとのタイトル数の分布

そのうち、ラスト・ワン・ツーにあたるタイトルは約32万冊であった。タイトル数としては、ロングテールの傾向を示しており、所蔵図書館が少ない資料ほど、タイトル数は多くなる。一方で11館程度に所蔵されている資料が冊数ベース（つまり図書館での専有スペース）ではボリュームゾーンとなることがわかった。なお、この数値は速報値であり、現在データ精度の検証を、多摩デポと進めている。

データを活用するための仕組み

現在、図書館が除籍する際の作業フロ

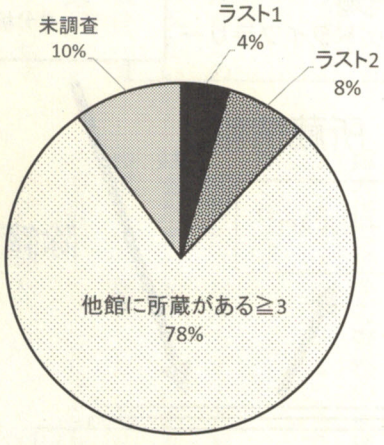

図2　除籍リストとの突合結果

ーの中に、このデータを取り込んでもらうために、どのようなサービスを提供できるか検討している。現状では、図書館が作った除籍したい本のリストを専用のウェブサイトに送ることで、ラスト・ワン・ツーかどうかのデータベースと突き合わせ、その調査結果のデータがダウンロードできる、という流れを想定している。

この自動化された突き合わせの結果の精度がどの程度のものなのか確かめるために、調布市立図書館の協力により、除籍候補リスト546件を使ってテストした。除籍リストにあったタイトルのうち、ラスト・ワン・ツーと判定されたタイトル数は全体の12%（65冊）であり、他館に4冊以上の所蔵があるものは78%を占める426冊という結果だった。

約10%はNDLやNIIに書誌がないことから、今回データを収集していない資料だった。こういった資料は随時調査対象として追加調査をしていくことを想定している。

多摩バーチャルデポジットライブラリーの可能性

これまでの取り組みをまとめてみよう。まず、所蔵情報は全国どこであっても、広域的に図書館に追加の負担なく収集できる仕組みができた。そして、図書館システムはそのアクセスに耐える体制がすでに整っている。データは適切に差分のみを更新していくことで低コストで維持できる。多摩デポとカーリルは現在、このデータを図書館で活用するための具体的サービスの開発を進めている。これは、ラスト・ワン・ツーのデータベースであると同時にすべての所蔵情報が入ったデー

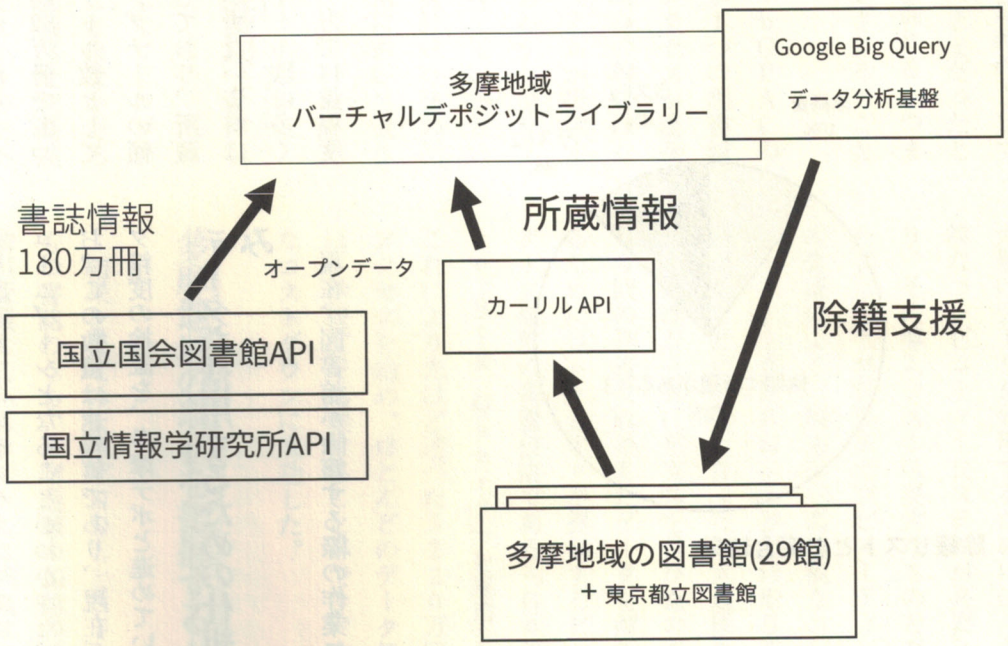

図3 多摩バーチャルデポジットライブラリーの仕組み

タベースだ。私達はこれを「多摩バーチャルデポジットライブラリー」と呼ぶことにした。

今回着手し、データ分析できたのはISBNが付与されている図書が対象である。1980年代以前のISBN付与が開始される以前の図書や、ローカルな郷土資料等の出版物については、今後の検討課題としている。これらの資料の同定には多くのノウハウが必要であることや、実物を確認する必要がでてくることから機械的な同定には限界がある。ただし、デジタル化が進むことにより効率化する仕組みをつくることができるだろう。マイクロボランティアなどの新しい考え方も有効である。

図書館の所蔵情報を広域的に分析し、さらに図書館の業務にも活かせることが理想である。そして、この情報はオープンであるべきだ。実際、今回このプロジェクトを一緒に進めている「NPO法人共同保存図書館・多摩」は市民による組織であって、行政そのものではない。しかし、定例ミーティングで重ねられた議論は、どれも公共的な視点に立ち、どうやったらよりよい状況を作り出せるかという建設的な議論であった。月に一度の定期ミーティングはとても楽しい時間となっている。

図書館をめぐっては、その役割や機能、あるいは効果や影響などについて様々な議論がある。

そうした議論を、〝感覚〟どまりではなく、より深いものにするために、様々な分析データを手軽に提供できる仕組みが実現できそうだ。

みんなで図書館サービスを担う

民間企業であるカーリルが、なぜこのようなプロジェクトに取り組むのかという質問を受けることがある。それは、カーリルが提供しているのは図書館サービスだという考えからきている。サービスを始めた当初、「図書館業界」というものが存在することをまったく知らなかった。自治体や大学が運営する図書館に対してサービスを提供している事業者と言えばいいのだろうか。それはシステム開発会社であり、業務委託業者であり、本棚メーカーであり、さらにいろいろな分野があるだろう。彼らの直接的な顧客は図書館であり、行政や大学となる。

一方でカーリルは、毎月50万人以上のユーザーに所蔵情報を提供している。図書館APIを通して間接的にサービスを利用しているユーザーも多い。ウェブサービスの規模としてはまだまだ小さいが、これはひとつの図書館サービスだ。だから、長期的な視野にたったとき図書館の提供できる資料の多様性を維持することは、カーリルにとっても、大きな課題と捉えている。

図書館におけるデータ解析の取り組みはまだまだ始まったばかりで、今回はその第一歩を報告した。これから多摩地域の図書館における共同保存をどうしていくのか、議論が深まることに期待している。この取り組みは、これまで多摩デポが築いてきた図書館との信頼関係の上に成り立っており、このプロジェクトに参加できたことに感謝している。