文字資料アーカイブズの現在

――特に検索可能性を中心に――

岡本真（アカデミック・リソース・ガイド株式会社）
mokamoto@arg-corp.jp

電子書籍ブームを踏まえつつ、改正著作権法の施行や国立国会図書館による大規模デジタル化の進展等、文字資料アーカイブズの現在を考察する上での前提となる近年の動向を概観した上で、特に文字資料の検索可能性の課題を論じる。

文字資料，改正著作権法，国立国会図書館，大規模デジタル化，検索

1. 文字資料アーカイブズの現在

本報告では、「情報処理技術は漢字文献からどのような情報を抽出できるか」という問いに対して、主に文字資料アーカイブズを中心に議論を展開したい。その際、これらの文字資料からの情報の発見・抽出にあたって重要となる検索技術の適用可能性を特に考えたい。

1.1 「電子書籍」元年

去る 2010 年は、電子書籍元年と喧伝された。実際、Kindle（アマゾン）や iPad（アップル）に続き、年末には GALAPAGOS（シャープ）や Reader（ソニー）が発売され、大きな話題となった。また、これらの機器・デバイスとして電子書籍だけでなく、ウェブサービスとして電子書籍のプラットフォームが相次いで提供もされている。

文字資料のデジタル化は、国文学研究資料館を中心に各種研究機関や研究者個人の手で行われて来ており、そこには相当程度の蓄積があるが、ここに来て、電子書籍元年の到来によって、画期的といっていい段階に入っている。権利処理を含め、これらのデジタルリソースの利用には、様々な課題はあるものの、「情報処理技術は漢字文献からどのような情報を抽出できるか」という問いを立てたとき、利用しうる対象データが爆発的に増加したことは喜ばしい。

1.2 改正著作権法

このように人力では扱い切れないほどの大規模なデジタルデータが文字の世界においても出現してきたが、文字資料アーカイブズの現在を語る上でさらに 2 つ重要な点が

ある。2010 年 1 月 1 日、改正著作権法が施行された。いわゆる検索エンジン合法化等が注目されがちだが、第 47 条 7 として、「情報解析のための複製」が認められたことを忘れてはいけない。議論の正確を期すため、条文を全文引いておこう。

（情報解析のための複製等）

第四十七条の七　著作物は、電子計算機による情報解析（多数の著作物その他の大量の情報から、当該情報を構成する言語、音、影像その他の要素に係る情報を抽出し、比較、分類その他の統計的な解析を行うことをいう。以下この条において同じ。）を行うことを目的とする場合には、必要と認められる限度において、記録媒体への記録又は翻案（これにより創作した二次的著作物の記録を含む。）を行うことができる。ただし、情報解析を行う者の用に供するために作成されたデータベースの著作物については、この限りでない。

これまでは、研究用途であっても文字データを大規模に収集・解析するには、一定の著作権処理が必要であった。たとえば、日本語コーパスの構築を目指して今年度まで 5 ヶ年計画で行われてきた特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」では、著作権処理に多大な労力を要したという[i]。法改正後、1 年を経た現在、まだ法改正の恩恵に関する目立った成果は見られないが、早晩この効果が見られるようになるだろう。

1.3 国立国会図書館による大規模デジタル化

文字資料アーカイブズの現在を語る上で、著作権法改正と並んで重要なのが、2010 年度の補正予算によって国立国会図書館が進める大規模デジタル化だ。同館の館長である長尾真が随所での講演等で語るところによれば、この事業に 100 億円を超える予算を組み、推定通りに進めば、1960 年代後半までに刊行された日本語書籍について、同館に所蔵されている限りデジタル化されるという[ii]。

ただし、一方の利害関係者である出版各社の意向もあり、この大規模デジタル化事業でデジタル化されるのは、あくまで版面の画像データとされている。その意味で、この事業の成果は、厳密には文字資料とは言えないのもまた事実である。しかし、これもあまり認知されていないようだが、国立国会図書館と一部の出版社[iii]の間では、「全文テキスト化実証実験」が昨年中盤から実施されている[iv]。この実験では、今年度中の結果のとりまとめを予定しており、その結果がもたらす影響、特に大規模デジタル化における手法を画像データから文字データへと進める効果をもたらすのか、大いに注目される。

2. 検索可能性という課題

さて、ここまで概観してきた「現在」を踏まえ、「情報処理技術は漢字文献からどのような情報を抽出できるか」という問いを考えていきたい。特に重要な論点と考えるの

は、「検索」である。

2.1 抽出・解析・蓄積の先にある課題

先に述べたように、著作権法の改正により、研究用途での大規模な情報の抽出や解析は極めて実現しやすくなった。この結果、本シンポジウムの開催趣旨にも掲げられている

i) 人手では不可能な大量のデータを扱い

ii) 人手による処理では帰納できない類の情報を抽出し、

iii) 得られた情報を機械可読かつ再加工可能な形式で蓄積する

という人文情報学の目的の最初の 2 点はこの先、飛躍的な進展を遂げる可能性が高い。しかし、3 点目の「得られた情報を機械可読かつ再加工可能な形式で蓄積する」はどうだろうか。蓄積された情報を利用するには、必要とする情報を引き出す仕組みが求められる。ここで出てくるのが、過去 10 年ほどの間に飛躍的に重要性が高まった検索技術であろう。確かに既存の検索技術、たとえば文の類似度を判定し、同一の、あるいは類似度の高い文を検索する技術は相当程度に確立されている。また、類似性とは異なる人間的な連想という思考方法を機械的な検索に適用するいわゆる連想検索の技術も徐々に普及してきている[v]。

しかし、世界の最先端を行く Google の検索技術をしても万人を納得させるには至っていない。なぜか。たとえば、類似度が同一の文と文の間、語と語の間でその優劣を判定する技術が確立されていないからだ。Google がまさに実践しているように、ウェブ上の情報は無数のウェブページ間の関係性をリンクと被リンクの関係性を抽出・計算することで、いわゆる集合知に基づく優劣関係の判定を行えてはいる。だが、要するに人気を指標とするこの方式（ペイジランク）の限界はつとに指摘されているところだ[vi]。とはいえ、一つの確立された方式ではあるが、この技術はこれから予測される文字資料の大規模なデジタル化には必ずしも有効ではない。ウェブ情報と異なり、データ間の関係性を必ずしも有していないためである。ここに文字資料アーカイブズが越えなくてはならない「検索」という課題がある。

2.2 「検索」を実現するための構造化

結局、課題は「検索」へと行き着く。一つの解として考えられるのは、ウェブ創発の初期から指摘され、最近も論議が高まっている文そのものに意味を持たせる手法、いわゆるメタデータ付与という手法であろう。この手法に関しては、たとえば国立国会図書館が 2010 年に発足させたデジタル情報資源ラウンドテーブル[vii]での議論や、同じく 2010 年に日本でも具体的なプロジェクトが始動した Linked-Open Data（LOD）[viii]といった動きがある。では、人文情報学の場合、どのような可能性が考えられるだろうか。

可能性の一つが、まさに本シンポジウムの最大の論点であり、開催趣旨で述べられている人文情報学の上述の 3 つの目的の先にある

> 最終的には、「××という特徴をもつ文献はどのような構造をしているのか」を機械可読な形式でモデルとして提出する、つまり「文献の構造の情報学的モデル」を確立する

だろう。実際にどのような手法が考えられるのかは、これからの議論によるが、ここで一つの提案をしておきたい。なお、これは「文献の構造」に直接的に関わることではなく、そのモデルが確立された暁に、必要とする情報を引き出す際のパラメーターの一つとして有用と思われるデータの整備に関わることである。

本稿の前半でふれたように、文字資料のデジタル化は大いに進展しつつある。特に近年の特徴は、原資料そのもののデジタル化であり、本文のアーカイブズ化である。しかし、いささか本文、データベースの区分で言えば、ファクトデータベースに傾斜しすぎていることを懸念する。言うまでもなく、人文学的な学問領域においては、特に先行研究の参照が強く求められる。いつ誰がどこでどのような形で、ある本文に言及しているのか、という情報を知らずして、文学や歴史学といった分野の研究は成り立たない。しかし、その割には、研究文献のデータベース、いわゆるレファレンスデータベースの整備が滞ってはいないだろうか[ix]。

もちろん、単に書誌情報を集約したレファレンスデータベースでは、「文献の構造の情報学的モデル」の確立に寄与するところは少ない。しかし、どのような論文でどのような本文が論じられているのか、ということを起点に、本文と言うなればその解釈の関係をメタデータとして内包する書誌情報であればどうだろうか。モデル確立に寄与することはもとより、その先での必要とする情報への到達の容易さ、つまり検索可能性にも大きく影響するだろう。本文だけでなく、この方面の研究にも力が注がれるよう期待を表明して、本稿を終えたい。

---

[i] 2010 年 3 月 10 日に開催された情報処理学会創立 50 周年記念（第 72 回）全国大会におけるシンポジウム「改正著作権法と IT」での前川喜久雄の発言等。http://www.ipsj.or.jp/10jigyo/taikai/72kai/event/39.html

[ii] 岡本真・仲俣暁生編著『ブックビジネス 2.0』（実業之日本社、2010 年）所収の長尾論文ほかを参照。

[iii] 「国立国会図書館における全文テキスト化実証実験の出版社等との共同実施について」＜http://www.ndl.go.jp/jp/aboutus/digitization_fulltext.html＞ によれば、2010 年 10 月 12 日時点で以下の各社である。

アーバンプロ出版センター、暁印刷、旭印刷、有田・海南のフリーペーパー Arikaina、岩波書店、イングカワモト、大月書店、快晴堂、紀伊國屋書店、共同印刷、語研、実

業之日本社、実務教育出版、渋沢栄一記念財団、寿限無、小学館、新人物往来社、新潮社、スタイルノート、青弓社、第一法規株式会社、第三書館、大修館書店、大日本印刷、太郎次郎社エディタス、筑摩書房、中央公論新社、東京創元社、東京大学出版会、東京電機大学出版局、読書工房、トランスビュー、日外アソシエーツ、パンローリング、フライの雑誌社、文藝春秋、ポット出版、まむかいブックスギャラリー、ミルグラフ

[iv] 国立国会図書館における全文テキスト化実証実験の出版社等との共同実施について、http://www.ndl.go.jp/jp/aboutus/digitization_fulltext.html

[v] 主に国立情報学研究所（NII）で開発が進められている連想検索エンジン「GETA＜http://geta.ex.nii.ac.jp/＞のほか、東京大学発のベンチャー企業であるプリファードインフラストラクチャーによる reflexa ＜http://labs.preferred.jp/reflexa/＞等がある。

[vi] たとえば、情報通信研究機構（NICT）の委託研究として京都大学の田中克己研究室を中心とする「電気通信サービスにおける情報信憑性検証技術に関する研究開発」の「Web コンテンツ分析技術」＜http://www.dl.kuis.kyoto-u.ac.jp/i-believe/＞で研究・開発が進められている。

[vii] デジタル情報資源ラウンドテーブル、http://www.ndl.go.jp/jp/aboutus/roundtable.html

[viii] たとえば、国立情報学研究所（NII）の武田英明らによる研究グループによって、LODAC Projects＜http://lod.ac/projects/＞が開始されている。

[ix] 詳しくは、岡本真「日本史研究におけるインターネットの学術利用－これまでの成果と、これからの課題」（『日本歴史』740、吉川弘文館、2010 年 1 月）を参照。