データ駆動による課題解決型人文学の創成

①　計画の概要

本計画は、従来は経験則に基づき進められ、データの蓄積という概念の無かった人文学分野の研究を、自然科学を含む他の分野にも共通し得る方法としてのデータ駆動型に再構築し、持続可能な社会を構築するためのデータインフラストラクチャーを人文学分野に築き、その利活用を通して多分野と協働し得る課題解決型の人文学研究を創成するものである。

日本語の歴史的典籍には、世界的にも稀な連続性を有する千年以上に及ぶ歴史的データが保存されている。歴史的データには気候変動や災害を含む様々な地球環境史の記憶から、多文化共生への知恵、心の問題や社会の在り方に関する考えといった人間社会の形成に関わる記述等に至る様々な記録が含まれるが、それらは紙というフィジカルデータを媒体とするため、そこに記録された情報の包括的な利用や分析は困難であった。そのため、人文学の専門的研究者であっても、環境の変動、社会の変化、人間形成の課題といった現代社会の直面する様々な課題の解決に積極的に参画することはなかった。現在進行中の大規模学術フロンティア促進事業「日本語の歴史的典籍の国際共同研究ネットワーク構築計画」（2014-2023）では、30 万点に及ぶ歴史的典籍画像が整備されつつある。本計画では、そのビッグデータを基盤として、範囲を明治時代にまで拡張し、人文学の分野に閉じられていた歴史的典籍に記録された情報やそのマテリアル情報等を広く自然科学、社会科学の研究者にも開き、歴史的典籍を軸とするデータ駆動型の人文学の研究環境を整備・運営してゆくものである。歴史的典籍データを機械可読型に整備し、自然科学・社会科学分野といった他分野の研究者との共同研究の成果をデータインフラストラクチャーに蓄積してゆく循環型の仕組みを構築し、人文科学の研究者が多分野と協働して自律的に現代社会にある様々な課題解決に十全に寄与する課題解決型の人文学を創成する。

②　学術的な意義

「日本語の歴史的典籍の国際共同研究ネットワーク構築計画」によって構築される大規模画像データベースにより、千年以上にわたって蓄積されてきた歴史的データの網羅的な閲覧が可能となったが、今回の計画ではその機械可読化と検索の高度化を推進し、マテリアル解析等の情報等を付加してデータを高度化することにより、非日本語圏や社会科学・自然科学分野に及ぶ広範な活用を可能とする。本研究では、大規模画像データベースの構築と並行した実証試験によって開発への糸口を得た AI の活用による認識技術を基盤として、その技術革新とともに得られたデータを人文学研究に提供し、データの分析成果としての補正値を技術開発に戻すという循環型の研究と開発を行うことを通して、人文学研究を新しい段階に推し進める。

例えば、AI の活用による認識技術の確立のためには膨大な学習データが求められるが、漢字文化圏における文字表記の変遷とその統計的分析を行う人文学の研究成果を認識技術の開発に提供し、認識技術の高度化の途上で得られた分析データを人文学に戻すことで、ビッグデータを用いた多角的な分析が可能となる。また、従来、書物はテキストを保存するものと考えられてきたが、マテリアルとしての書物には植物の DNA が保存されており、人間が関わることで毛髪、手垢等の人間由来の成分も保存されている。マテリアルとしての書物の情報は、注目されることが少なかったが、人間史の分析データともなる。加えて、こうしたデータがいずれも千年以上に及ぶ連続性を有して日本という国に残存することは世界的にも希な現象であり、地球環境史と人間と社会の営みを解明し、現代社会の直面する様々な課題を解決するための世界規模のエビデンスデータとなる。ビッグデータを用いたデータ駆動型の人文学研究の創成は、世界に先駆けて人文学研究のパラダイムシフトを行うものとして大きな意義を有している。

③　国内外の動向と当該研究計画の位置づけ

今世紀初頭頃からイギリス、フランスを中心としたＥＵ諸国及び米国、韓国、中国などで、それぞれの国や地域に蓄積された書物や文書類のデジタル化とその公開が国家的事業として急速に進められてきた。人文学研究においても、デジタル・ヒューマニティーズに代表される、デジタル化された資料を用いた分析が先駆的研究として行われるようになり、そのための分析ツールなども広く公開されてきてはいるが、こうしたデータの利用と研究は、旧来型の人文学研究の範囲に留まっている。

本計画は、「日本語の歴史的典籍の国際共同研究ネットワーク構築計画」による大規模画像データベースの構築により、ようやくに先行する諸国に追いついた状況にある日本のデジタルリソースの活用を人文学に留まらない多分野に開き、蓄積されてきた歴史的典籍の活用の推進のためのデータインフラストラクチャーの構築、ビッグデータのコンテンツ解析・マテリアル分析への提供と得られたデータの環流と蓄積といった、人文学分野と社会科学・自然科学分野との循環的研究環境の構築を通して、データ駆動型の人文学を創成するという、世界の現状を鑑みても先駆的な計画であると言える。

④　実施機関と実施体制

本研究は、国文学研究資料館を実施主体とし、情報システム研究機構の協力をあおぎ、とくに大規模学術フロンティア促進事業「日本語の歴史的典籍の国際共同研究ネットワーク構築計画」において学術協定書及び覚書を交わしている下記の諸機関とともに遂行するものである。以下、実施内容毎に記す。調整中の機関についてはその旨を記した。

(1)データインフラストラクチャーの構築：国立国語研究所、国立情報学研究所、国立国会図書館、情報システム研究機構データサイエンス共同利用基盤施設、人文情報学研究所、公立はこだて未来大学、東京大学文学部、慶應義塾大学文学部、凸版印刷（株）

(2) コンテンツ解析からの展開：国立極地研究所、茨城大学地球変動適応科学研究機関、奈良文化財研究所（共同研究実施中、学術協定等については調整中）
(3) マテリアル分析・解析：立命館大学文学部、実践女子大学文芸資料研究所、奈良先端科学技術大学院大学向川研究室（共同研究実施中、学術協定等については調整中）、龍谷大学文学部（学術協定等については調整中）
(4) 人文系データ分析技術の開発：国立国語研究所、国立情報学研究所、公立はこだて未来大学、法政大学能楽研究所、ゲーテ大学フランクフルト・アム・マイン言語学・文化学・芸術学部（独国）、ハイデルベルク大学日本学研究所（独国）、カリフォルニア大学バークレー校東アジア図書館（米国）、フリーア・ギャラリー（米国）、大英図書館（英国）

事業を統括し推進する部署として、国文学研究資料館においてデータ駆動研究センター（仮）を設置し、マテリアルデータの蓄積を担当する実験ラボを設置する。また、外部の意見を反映させプロジェクトを企画・実施していくために、センターに各種委員会を設置するとともに、顧問及びアドバイザー制度を導入する。

⑤ 所要経費

本計画は、総額38億円を想定している。内訳は以下の通り。
委員会経費：5,000千円×10年　共同研究経費：100,000千円×10年　成果公開費：40,000千円×10年　データベースシステム経費：40,000千円×10年　センター人件費：140,000千円×10年　データ整備費：50,000千円×10年　管理経費：5,000千円×10年　総額、単年度380,000千円×10年

⑥ 年次計画

第1年時にデータ駆動研究センターを設置し、データ駆動型のシステム開発を立案・設置し、以下の研究計画を遂行する。
(1) データインフラストラクチャーの構築
データ駆動型システムの開発（第1～2年時）　データアーカイブ機能の強化：多分野での利活用を想定した権利関係等のガイドラインの検討・策定と分野横断的データカタログの整備（第1～2年時）　海外発信・連携機能の強化：国内外でのシンポジウム開催（第1～10年時、シンポジウムの開催）　データ間の時系列等接続関係の整備（第2～3年時）
(2) コンテンツ解析からの展開
典籍防災学の拡大（第1～3年時、第4～6年時）　典籍人類学の構築（第1～3年時、第4～6年時）
(3) マテリアル分析・解析
マテリアルとしての書物の分析技術の確立（準備研究：第1～2年時、本研究：第3～5年次、第6～8年次）　マテリアルとしての書物の復元技術の確立（同上）
(4) 人文学系データ分析技術の開発
他分野からのメタデータ付与に関する合意形成及び汎用的仕組みの検討と開発（第1～2年時）　画像検索・解析技術の精度向上と可視的把握技術の確立（学習データ作成：第1～2年時、本研究：第3～5年時、第6～8年時）　AI技術に基づく機械可読データの自動化の開発（同上）　国際テキスト（TEI）へのフォーマット作成及び作成ツールの開発（第1～3年時）　画像・テキスト等人文系資源に基づくデータ駆動型人文学研究への展開（第5～10年時）
センターには人文学に限らない若手研究者を配置する。共同研究では若手研究者の比率を20%以上に設定し、データ駆動型の人文学研究に精通した研究者を育成するとともに、研究と社会を結ぶ高度な専門知識を保持するインタープリタ等を育成し、研究成果の社会還元・普及にも努める。また、本事業終了後もデータのリプレイス等を継続してゆく。

⑦ 社会的価値

日本語の歴史的典籍には、我が国に生きた人々の思想と感情の歴史のみならず、対外交流史、農林水産業・工業を含む産業史、災害史、気候変動を含む自然史といったさまざまな分野に関わる事柄が記録されており、千年を越える時代にわたって同一の言語環境における連続したデータを取得することが可能であることは世界的にも稀なことである。こうした典籍は、従来は主として歴史的典籍に関わる専門家にのみ活用されてきたが、今回の研究で画像データを機械可読化し活用することにより、多分野における利活用が可能となる。人文学の研究者が他分野の研究者と協力することにより、日本語の歴史的典籍に記録された情報が知的価値に加えて経済的・産業的価値を持つデータへと転換し、SDGsの示す、「11 住み続けられるまちづくり」、「13 気候変動に具体的な対策」といった目標にも貢献することができる。

⑧ 本計画に関する連絡先

ロバート キャンベル（大学共同利用機関法人人間文化研究機構国文学研究資料館）