No.59 （グランドビジョン⑧）

人文知を基盤としたAI技術の応用による真の無障壁社会の実現

① ビジョンの概要

進展著しいAI技術とデジタル人文学の総合的利用により情報（特に文化的情報）へのアクセスを飛躍的に高め、全ての人にとっての無障壁社会を実現する。距離や年齢、障害、言語による障壁をなくして、文化多様性の復元、社会問題の解消、地域の活性化を図り、文化的に豊かで生きやすい社会を作る。

② ビジョンの内容

今日、高度な情報化により膨大な情報へのアクセスが可能になり、一見豊かな生活が実現したかに見える。しかし、これらの情報化は特定の産業や用途に偏っており、いまだそのアクセス方法も複雑で、利用は特定の人に限定されたものとなっている。地方には豊かな言語、文化・伝統が残っているが、それらは現在保存されないまま消滅の危機に瀕している。多くの文化資料、地域文献資料は保存されていたとしても紙のままで資料館、図書館、博物館に収蔵され、方言や伝統文化はアナログ資料のまま収蔵庫奥深く保管されている。こうした過去の膨大な記録へのアクセスには制限が多く、ほとんど活用されないままで眠っている。

一方で、情報のバリアを安易に技術的に取り除こうとする動きが却って、ディジタルディバイドを作り、地域格差と地域文化の消滅による画一化や情報格差を作り出している側面もある。この格差が社会的な弱者を生み、地域文化の衰退・文化の画一化を促進しつつある。

このような中、技術的に手に入れやすい情報だけに目を向けさせるのではなく、人文学が蓄積してきた知識と技術をAI技術とともに創造的に活用することで、情報バリアを解消した真の無障壁社会を作り、人々を文化的に豊かに生きやすくし、多様な文化に活力を取り戻していくことが求められる。このような環境の整備は、たんに工学的な技術提供を受けるだけでは実現せず、対象となる資料に対する深い知見と情熱をもって人文学者が積極的に参加して構築する利用環境によって初めて可能になるものである。

③ 学術研究構想の名称

言語データ解析による人文知を基盤としたアーカイビングシステムの構築：「人文知コンシェルジュ」サービスの実装

④ 学術研究構想の概要

近年のAI・言語処理技術の急速な発展により多くの要素技術がすでに開発されている。しかし、十分な研究や実用化が進んでいない応用領域が数多く存在する。自動音声・文字認識、単語・統語・意味解析、機械翻訳などは現代の大言語では実用化が進んでいるが、各地の方言・危機言語や、数十年前の資料、高齢者・子供・障害者の言葉においては全く不十分な状態である。こうしたマイナーな言語データは、短期的には経済的価値を生むことはなくとも、長期的にはかけがえのない価値を持つものである。価値を最大限に活かして応用するためには、過去から積み重ねられてきた人文学的な英知を欠かすことができない。

本研究構想では、人文学的専門知識・技術をAI・自然言語処理技術と融合させることで、これまでに手が付けられていない多くの情報を機械可読・アクセス可能にし、利用しやすい環境を整備することによって情報バリアを取り除くことを目的とする。そのために、以下の3つの課題を立てて研究を推進する。

・蓄積の少ない言語資料のデジタル化・共有化
・図書館・博物館・資料館・人文系の諸データの統合
・構築した資源にもとづく研究および社会への応用の推進

これらの取り組みを通じて、デジタルデータとオープンサイエンスに依拠した新時代の人文学の推進を図る。同時に、一般市民に向けて文化資源を分かりやすい形で提供し、価値の一元化による差別を解消し、文化的に豊かで生きやすい社会作りに貢献することを目指す。

⑤ 学術的な意義

本研究構想では、蓄積されたデジタル資料の利用価値を大きく発展させるために、文献・画像・音声・動画資料に含まれる多様な日本語のデータを解析し、単語として意味を扱えるように整備する。また、資料中の固有表現を抽出し、固有表現辞典、時空間情報、さらには博物館資料の情報などの人文学知識との関連付けを行う。こうしてデジタル資料と人文知を融合させた、コンテンツの全文検索が可能なアーカイビングシステムを構築する。そのために、方言音声の自動テキスト書き起こし・形態素解析・辞書見出し化・標準語

訳を行い、方言のアーカイブ構築を容易に行えるシステムを開発する。また、図書館・資料館等の前近代から明治期にかけての古文・文語文を解析し、現代語に機械翻訳を行う技術を開発する。

こうして構築するデジタルデータの利用環境は、研究者にとってはオープンサイエンスを基礎とした人文学 DX の実践の基盤となるだけでなく、人文学内部での垣根を越えた融合研究にも資するものとなる。

⑥　国内外の研究動向と当該構想の位置付け

国内では、図書や音声・映像などの文化資源の電子化は、長い間デジタル化とメタ情報の整備に留まっており、データの中身に踏み込んだ高度な利用を行うことができない状況が続いてきた。しかし、近年になって深層学習技術の発達により、これらのデジタル情報の中身の解析・活用が重要になってきている。本研究構想は、こうしたマイナーでありながら図書館や博物館資料の多くを占める資源を対象として、言語解析技術を開発して飛躍的にアクセシビリティを向上させる。ヨーロッパでは、Europeana や Gallica などこうした環境整備が進んでおり、日本語の資料でもこれらに伍する環境を整備することが求められている。

⑦　社会的価値

過去の文化資源へのアクセスを飛躍的に向上させることで、国民の知的好奇心に答えられる利用しやすいアーカイブを整備し、豊かな文化にふれられるよう貢献する。これまで図書館・博物館・研究所等で別々に蓄えられてきた言葉・地理情報・歴史年代・モノ資料の知識を相互に結びつけることにより、各機関が保有する資料の利用価値を高めることが期待できる。

⑧　実施計画等について

１～２年目に、コンテンツの全文検索が可能なアーカイビング環境を設計、辞書データの整備を実施。方言音声書き起こし、古典籍や近代資料の解析、固有表現抽出を実施する。その後、５年目までに、デジタルミュージアムの機能を大幅に拡張、固有表現データをもとに情報を相互にリンクさせた次世代の利用環境の設計・開発を行う。５～７年目に、方言音声データ（文化庁緊急調査）の解析を完了。前近代のテキストの解析を実施。国語辞典の見出し語、意味分類、地図上の位置、年表、人名などから、各種アーカイブ上の原情報にアクセス可能にする。また、機械翻訳によって古文や方言を現代標準語訳する。最終年度までに、研究者だけでなく国民だれもが使いやすい文化情報アーカイブを「人文知コンシェルジュ」サービスとして実装する。

実施機関と実施体制としては、国立国語研究所が中心となり、人間文化研究機構が実施するデジタルヒューマニティーズ計画を下支えし、新しい人文学の研究環境を整える。そのために国語研内に「言語アーカイブ解析センター（仮称）」を置く。総経費 1,957,600 千円

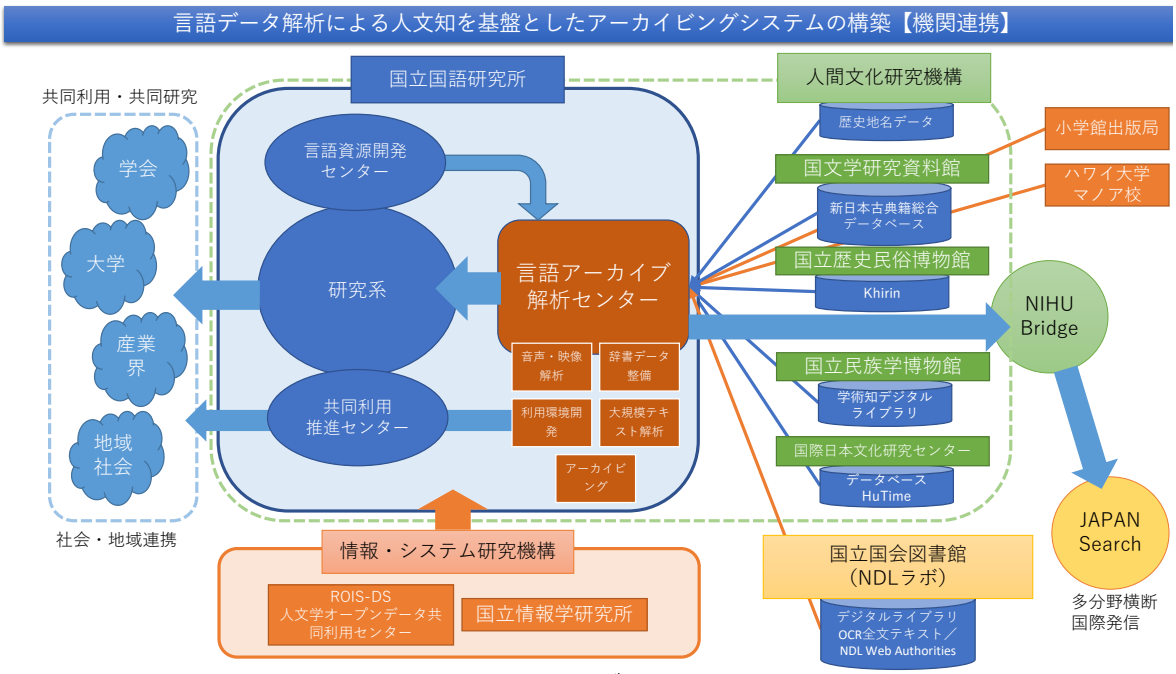

図２　設置する言語アーカイブ解析センターと機関連携

⑨　連絡先　前川　喜久雄（国立国語研究所）