歴史ビッグデータ：構造化ギャップを克服するワークフローの構築と過去世界の統合解析

Historical big data: integrated analysis of the past world by the workflow that bridges data structuring gaps

*北本 朝展[1,2,3]、市野 美夏[1,2]
*Asanobu Kitamoto[1,2,3], Mika Ichino[1,2]

1. 国立情報学研究所、2. ROIS-DS人文学オープンデータ共同利用センター、3. 総合研究大学院大学
1. National Institute of Informatics, 2. ROIS-DS Center for Open Data in the Humanities, 3. SOKENDAI

1. はじめに

過去の書籍や文書から情報を抽出し、それを統合することで、過去の世界を復元して分析する。このような歴史研究を我々は「歴史ビッグデータ」と呼ぶ。それはこのアプローチが、現代を対象に行われる「ビッグデータ」の研究と同じ構造や同じ目的を持つため、現代ビッグデータを過去に延長していくことに意味があると考えるからである。しかしそこに立ちはだかるのがデータ構造化である。過去のドキュメントにくずし字（手書き文字）で書かれたデータから、過去の世界を計算処理で復元するための高品質データを整備するには、デジタル化から品質管理に至る長大なデータ構造化ワークフローを支援する情報基盤が必要である。しかし自動的な構造化は困難なため、人間と機械の共同作業による効率的なデータ構造化ワークフローが必要である。そこで本研究は、非構造化データ（画像・テキスト）を構造化データ（解析準備データ）に変換するワークフローを設計し、再利用かつ検証可能な人文学データセットを構築するための情報基盤を構築する。

EUでは「Time Machine Flagship」（https://timemachineproject.eu/）という巨大プロジェクトが200機関以上の参加を得て立ち上がりつつある。そして、イタリアのベニスやオランダのアムステルダムなど、都市の歴史のビッグデータを集めて時空間を自由に行き来する「タイムマシン」の構築が始まっている。この動向は日本やアジアにはまだ波及していないため、本研究で構築する情報基盤は日本の拠点となってグローバルな活動と連携できる可能性がある。

2. 課題

過去の世界を復元するための研究はこれまでも数多く行われてきたが、歴史ビッグデータ研究は以下の点で既存のアプローチとは異なる。

第一に、対象とするデータの種類である。例えば古気候研究の場合、気候に関するあらゆるデータを用いるため、書籍や文書に限らず、自然界に残された痕跡（アイスコアや年輪などのプロキシデータ）なども活用することになる。しかし歴史ビッグデータの対象はあくまで人間が残した記録に限定し、文字記録の読み解きを含めた新しいデータ構造化の研究に焦点を合わせる。

第二に、対象とする分野である。単一分野の研究では、過去の世界の一部の現象のみを対象とし、それ以外の現象には注意を払わないことが多かった。例えば同じ日記を研究対象としていても、書誌データや抽出データは研究者グループを越えて共有されず、多数の研究者が同じような作業を繰り返す状況に陥ることが多かった。この状況を解決するため、歴史ビッグデータは分野横断的に活用可能な構造化ワークフローを提供し、情報共有のメリットを活用した研究を可能とする。

3. データの掛け合わせと読み替え

現代ビッグデータを過去に延長するには、技術の過去への延長に加えて、コンセプトや方法論の延長も重要な課題である。

第一に、データの掛け合わせとは、異なるデータを重ねて意外な関係性を見出すという方法論である。その典型的な例が地図である。複数のデータを位置合わせして重畳表示することで、データから得た洞察をアクションにつなげることができる。そこで課題となるのが、APIの相互運用性や語彙の共有などである。この問題を解決するために、我々は研究グループの今後の研究課題を共有し、作業の重複を避けてお互いの強みを活かすことで、限られたリソースを最大限に活用した情報基盤を開発している。

第二に、データの読み替えとは、ある目的に作られたデータを別の目的に再利用することの価値を見出す方法論である。現代ビッグデータにおいて有名な例は、車の走行データを震災時の通れる道マップに再利用するという事例であるが、同様のアイデアは歴史ビッグデータでも有用なはずはずである。例えば、人名録の変遷は気候変動の社会影響評価に使えないかなど、柔軟に発想を巡らせてデータを創造的に活用する必要がある。

4. 同床異夢を越えて

歴史ビッグデータ研究は、多分野を融合した研究である。もちろん人文学と理工学など文理の間には大きな違いがあるが、理工学の中でも分野による考え方の違いは決して小さくはない。こうした違いをどのように乗り越えるか。我々の基本的な考え方は、まず同床異夢であること、すなわち共同研究のメンバーが目指す個々の夢は異なることを認めた上で、なお同床であることの意義を積極的に評価するというものである。例えばデータやツールは夢が異なるものの間でも共有できるはずである。こうした共有のメリットを最大化した上で、個々の研究者は過去世界の異なる部分の復元に挑むというのが歴史ビッグデータ研究の構想である。

キーワード：歴史ビッグデータ、古気候、データ構造化、情報基盤、異分野融合研究、歴史的文書
Keywords: Historical big data, Paleoclimate, Data structuring, Information infrastructure, Multi-disciplinary research , Historical documents

