Vol.2019-CH-120 No.13
2019/5/11

「人文科学とコンピュータ分野」における研究資源と情報技術を考える

後藤真[†1]　阪田真己子[†2]　松村敦[†3]　山田太造[†4]

概要：これまで本研究会では，文学・歴史・考古・音楽・舞踊などの人文・芸術等の研究におけるデータを対象に，検索・分類・提示・発見などを実現するための情報技術の方法論・適用手法等について多数の報告がなされてきた．CH120 では，ここでのデータおよび方法論について議論し，本研究会の特徴を見いだし，本研究会の将来の方向性について議論していく討論会を開催する．

キーワード：企画セッション　人文学の研究資源　CH の情報技術

MAKOTO GOTO[†1]　MAMIKO SAKATA[†2]
ATSUSHI MATSUMURA[†3]　TAIZO YAMADA[†4]

Keywords: Special Session, Resource for Humanities, Information technology for SIG-CH

1. はじめに

本セッションでは、「人文科学とコンピュータ分野」における研究資源と情報技術を考えることをテーマとしてすえた。本研究会が進めてきた人文科学へのコンピュータ利活用（CH）の分野においては、多くの場合、研究資源となる人文学資料に対し、分析手法として情報技術を用いる研究の枠組みが用いられてきた。それは、人文学の問題を解決するためにこれまでとは異なる手法を導入する点において、また情報技術の新たな応用先を見つけるという点において、重要な枠組みであった。そして、今後も基礎的な枠組みであると思われる。

図 1 は、筆者の一人である後藤が考える一般的人文学の研究プロセスとそれに関わる CH 研究の位置付けである。以下のような流れが考えられる。

1．まず対象となる基礎的な資料を発見する。2．その上でそれらの資料を分析する。3．分析結果を論文とし、必要に応じてアウトリーチ等を行う。

この一連の流れの中で CH の研究が生まれる。後藤は、これらを発見系・解析系・可視化系という大きな三つの枠組みとして整理することで、研究傾向をみる考え方をとっている。無論、すべての研究がこの三つに単純に分けられるわけではなく、複数の系統の要素を持つものや、この三つの枠組みには当てはまらないものも当然ある。しかし、まずは議論のスタートラインとして、このような整理を行い、この中で資料や技法はどのように考えられるのかを考えてみたい。

この枠組みを切り口として、CH 研究会のこれまでの研究発表の傾向などから改めて整理することで、どのような資源や技法があるかを再確認し、さらに高度な研究へ進む材料とできるのではないだろうか。

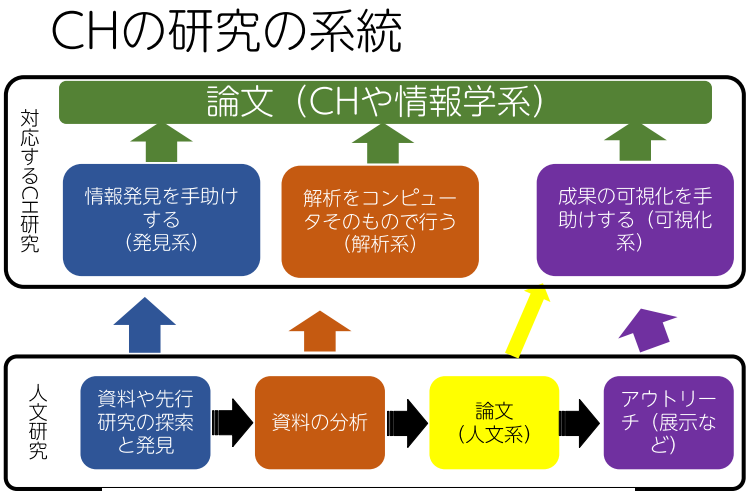

図 1　CH の研究の大まかな系統

2. CH における研究資源

まずは研究資源である。CH 研究会における基礎的な研究資源は、1 章で述べたように、当然人文学の資料が中心となる（情報技術そのものを資料とするわけではない）であろう。しかし、この資料も様々である。図 1 の流れに沿うならば、左の二つ（発見・解析）については多くの場合、人文書の資料そのもの（例えば古文書や古典籍、音楽や考古資料、舞踊など）が対象となる。一方で、右の二つ（可視化系）であれば、研究論文そのものや、成果のより抽象

†1 国立歴史民俗博物館
National Museum of Japanese History
†2 同志社大学
Doshisha University
†3 筑波大学
University of Tsukuba
†4 東京大学
The University of Tokyo

的なものが資料となりうる。可視化系の場合は前者に比べ、より抽象度が高いファクトデータを対象とすることになる。この場合はあまり分野に関係なく、いかに研究者の思考を抽出するかに重点が置かれることになる。

一方で、前二者（発見・解析系）の資料については、人文学内の分野ごとにも特性が出ることになり、その特性に応じた手法を選択することになると思われる。CH 研究会ではこれまでに下記のような分野が対象となってきた（順序はおおむね NDC 分類順）。

宗教学　特に仏教典籍については、SAT（大正新脩大蔵経テキストデータベース）の存在が大きく、テキスト・画像などの解析を中心に様々な手法の研究が行われてきた。聖書研究などは DH の基礎の基礎でもある。基本的には、教典そのものの研究が多い傾向があり、宗教史関連資料はそれと比較すると多くはない。

歴史学　日本史学・東洋史学（とりわけ中国史学）が比較多数を占める。大規模データベースの開発とともに、テキスト解析なども行われる。また、漢字に関わる研究も資料という文脈においてはこの分野に当てはまる部分もある。主に「文字」「文」を対象とする。なお、歴史人口学などの数的処理も行われてきた。写真を除けば、前近代の資料を対象とするものが多く、発見系の研究が多い。

民俗学・文化人類学　フィールドノートの解析や動画アーカイブなどが主たる研究となってきた。可視化やよりメタな方向で議論を進める志向が強いといえるであろう。なお、やや広げて地域研究とすると、データ発見系の研究が一挙に増えてくる。対象地域は日本とアジアが中心であり、アフリカやラテンアメリカを対象としたものは少ない。

舞踊　CH 研究会においてはモーションキャプチャと不可分といっても良い関係となっている。一部で動画等の解析も行われている。幅広いが海外の庶民の舞踊などはあまり多くない。

言語　国語学に関わる研究は、日本史と関わる漢字や訓点・コーパス開発などテキストの開発とともに進められてきた側面が大きい。音声研究は多くない。

文学　計量文献学的な研究は DH との関係も相まって隆盛である。圧倒的にテキストを対象とする研究が中心となっている。文学資料を使ったくずし字翻刻なども近年行われている一方で、非漢字圏かつ非英語圏の文学研究は多くない。

なお、近年新たに学会ができるなどの側面もあり、あまり対象とされていない分野として、考古学（考古遺物や遺跡などが対象となり、モノの系統樹分析や、古墳の特徴などの形状解析がある）、教育学（特に教科書を含むテキスト解析）などがある。しかし、これらは今後も CH 研究会としても対象となり得る分野である。

3. CH における情報技術

次に、関連する情報技術について触れる。こちらは発見・解析・可視化の順に触れていきたい。

発見系研究　一時期はメタデータの標準化と応用が重要な位置を占めていた。大型データベースである nihuINT に関わる研究を中心に、どのようにデータベースを統合的に検索可能とするかといった観点が大きかった。これらの研究は、近年では Linked (Open) Data や Semantic Web などの技法に変わりつつある。手法ではないが、基盤整備や辞書・資料のインタフェース改善も本系統に繋がりうる。画像の発見と共有という観点からは、IIIF は近年の重要な潮流となっている。また、クラウドソーシングのような発見と「人にいかに手伝ってもらうかのアーキテクチャ」を含み込んだ観点からは、発見と解析の両者を含み込んだものも存在する。一方で、音声や動画の解析による情報発見などはあまり多くない。

解析系研究　テキストの計量解析は極めて重要な位置を占める。語彙の頻度解析や共起関係の分析により、テキストの特性をどのように理解するかの研究は、先述の通り DH の研究とも関係して特に研究が分厚くなってきた。また LDA によるテキストの類似度分析などは発見系への発展も含みつつ検討が進められてきた。また TEI もデータ構築だけではなく、分析の視点を含む系統として、解析系にあてはまるといえる。そして舞踊の解析などのモーションキャプチャも技法としては重要な位置を占めているであろう。また、近年の重要な動向であるディープラーニング等を用いた研究では、くずし字の自動翻刻などは情報発見の基礎ツールとなるとともに、それ自体が画像解析の手法となっている。小袖模様のディープラーニングを応用した解析など、画像解析でのディープラーニング活用が多い傾向がある。

可視化系研究　一時期は GIS が極めて大きな位置を占めていた。研究成果を地図上に落とし込み、それをいかにわかりやすく示すかといった研究が多数行われ、広く市民権を得た。一方で 3D モデルなどによる可視化などは、モーションキャプチャと関連するものなどはあるものの、これまでの研究の中では多いとは言えない。また AR、VR に関わる研究も CH 研究会の中では大きな潮流とは言えないのではないだろうか。

4. おわりに

本予稿では、あくまでも代表的なものを列挙したに過ぎない。1 章でも述べたように、このような分類に当てはまらない研究も存在する。例えば博物館展示などを含むユーザインタフェースなどは、この枠組みとも異なる視点を持っていると思われる。セッションにおいて、多くの資料やその特徴、技法の傾向や時間的変遷などまで含めて議論できれば幸いである。