公共的な利活用に向けた人文学における研究データ構築の諸課題

永崎研宣
一般財団法人人文情報学研究所　主席研究員

学術フォーラム「デジタルデータ・社会調査データの公共的な利活用に向けて」）
2023 9月24日（日）13:00～16:00

自己紹介

- インド仏教学で筑波大大学院博士課程哲学・思想研究科単位取得退学
  - その後、関西大学で論文博士号（文化交渉学）
- デジタル技術の人文学への応用の研究は院生時代から：初めての発表は電子情報通信学会の研究会
- 東京外大ｱｼﾞｱ･ｱﾌﾘｶ言語文化研究所任期付き研究員⇒山口県立大学国際文化学部専任講師／助教授⇒研究所設立に参画
- 現在：
  - 一般財団法人人文情報学研究所　主席研究員
  - 国立国会図書館研究員（委嘱）
  - 国文学研究資料館客員教授
  - 東京文化財研究所客員研究員
  - 広島大学文学部客員教授
  - 東京大学大学院人文社会系研究科非常勤講師（人文情報学関連科目2コマ）
  - 沖縄県立芸術大学客員研究員
  - COB member, Alliance of Digital Humanities Organizations – 国際DH学会連合の各国代表委員会
  - Convener, SIG East Asian/Japanese, TEI Consortium - テキストデータ構築の国際デファクト標準ガイドライン
  - Liaison member, ISO/IEC JTC1/SC2 (SAT Committeeとして) – Unicode/文字コードの国際標準規格
  - Committee member, ISO/TC37/SC2 - 言語コードの国際標準規格
  - Task force member, Comprehensive Digitization and Discoverability Program, North American Coordinating Council on Japanese Library
  - 情報処理学会人文科学とコンピュータ研究会運営員会委員
  - 日本学術会議7委員会合同デジタル時代における新しい人文・社会科学に関する分科会委員長

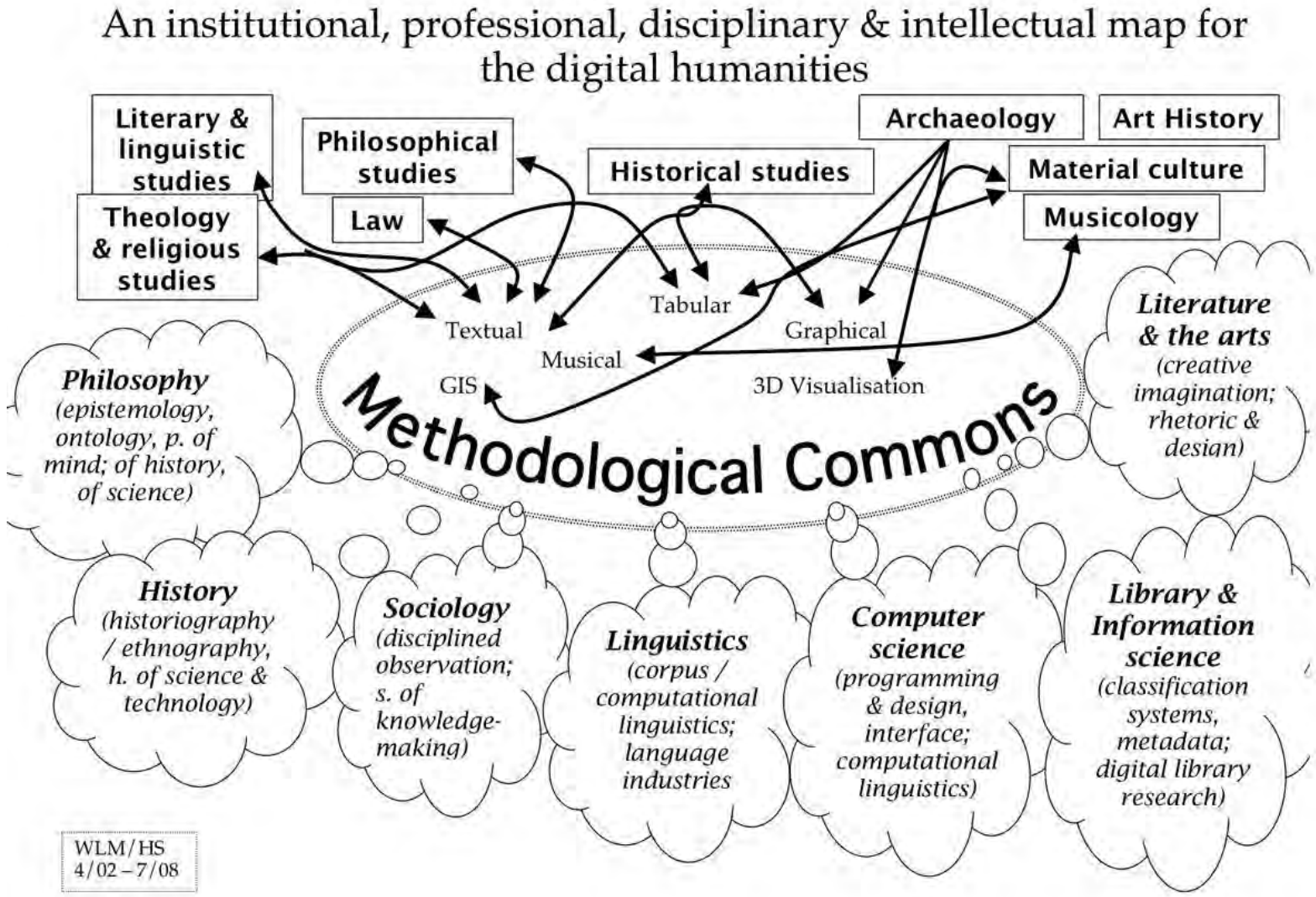
研究データを扱う人文学の国際的な潮流としての**デジタル・ヒューマニティーズ(DH)**

- ADHO (国際デジタル・ヒューマニティーズ学会連合)を中心に展開
  - 2005年頃に欧米の学会が連携して設立／2006年に第一回大会

- 2006年頃からの欧米における政策的な展開
  - National Endowment for the Humanities (米国) にてOffice of DH設立に向けた議論開始（⇒2008年設立）
  - DARIAH（人文学向けデジタルインフラ事業）の計画が始動
    - ESFRI Roadmap (研究インフラに関する欧州戦略フォーラム)の俎上に
    - ⇒2008年に準備フェーズ開始
    - ⇒2014年にERICにて正式に設立
  - ⇒グローバル・サウスとの研究協働にも取組み

様々な研究支援によりデータの構築とツールの開発／オープン化による共有

デジタル人文学の理念的背景



方法論の共有地

共通の場としてのデジタル技術を介して人文学・情報学分野の多様なディシプリン同士に対話と新たな可能性を生み出す。

https://eadh.org/methodologies
Willard McCarty and Harold Short, METHODOLOGIES, Pisa, April 2002

The Digital Humanities Manifesto

- 2008-2009年、主に米国の研究者が中心となって作成された。
  - https://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf
- 「オープン」が謳われている。

「公共／Public」と人文学

- Public (Digital) Humanities
  - https://4humanities.org/
    - An advocacy initiative for the humanities focused on placing the value of the humanities before the public.
  - Transcribe Bentham（英国）／みんなで翻刻（日本）
    - デジタル文字起こしをクラウドソーシングで⇒公共的なデータ構築へ
- Public History
  - パブリック・ヒストリー入門―開かれた歴史学への挑戦
    - 勉誠出版, 2019年
  - パブリックヒストリー研究会
  - 学術雑誌『パブリック・ヒストリー』
    - 欧文タイトル：Journal of History for the Public

人文学の「研究データ」とは？

- アナログ資料からのデジタル画像やテキストデータ
  - ⇒「**デジタルアーカイブ**」の画像や文字起こしテキストデータ
    - ＝人の手／判断をあまり入れずに二値化したデータ
  - 研究の効率化という観点で有用性がある
  - ＝「研究**の役に立つ**データ」
    - これを志向する研究開発が盛んに行われている
  - この種のデータと研究成果との間には**かなりの隔たり**がある

- 「研究成果の直接的な基礎となったデータ」としての研究データ
  - 「ノートやメモをとって検討した」
  - 「テキストを読んだ印象から判断した」
  - 「触ったら感触でわかった」
  - ⇒基礎資料に対するアノテーション／注記のようなもの

人文学の「研究の役に立つデータ」の例

- 全般：国立国会図書館デジタルコレクション
  - 各地の「デジタルアーカイブ」
- 日本語学：中納言（各種日本語コーパスの検索・分析サイト）＠国立国語研究所
- 日本史学：東大史料編纂所・国立歴史民俗博物館等のデータベース
- 日本文学：国書データベース＠国文学研究資料館
- 仏教学：SAT大蔵経テキストデータベース
  - デモを少し： https://21dzk.l.u-tokyo.ac.jp/SAT/

人文学の「研究データ」とは？

- アナログ資料からのデジタル画像やテキストデータ
  - ⇒「**デジタルアーカイブ**」の画像や文字起こしテキストデータ
    - ＝人の手／判断をあまり入れずに二値化したデータ
  - 研究の効率化という観点で有用性がある
  - ＝「研究**の役に立つ**データ」
    - これを志向する研究開発が盛んに行われている
  - この種のデータと研究成果との間には**かなりの隔たり**がある

- 「研究成果の直接的な基礎となったデータ」としての研究データ
  - 「ノートやメモをとって検討した」
  - 「テキストを読んだ印象から判断した」
  - 「触ったら感触でわかった」
  - ⇒基礎資料に対するアノテーション／注記のようなもの

研究成果の直接的な基礎となったデータ

- 「ノートやメモをとって検討した」
  - ⇒フィールドノート
  - ⇒テキスト／画像／動画／3D等へのアノテーション
  - ⇒アノテーションを集約したノート
    - ⇒工学オントロジー的な体系化があり得る？

- 「テキストを読んだ印象から判断した」
- 「触ったら感触でわかった」

研究成果の直接的な基礎となったデータ

- 「ノートやメモをとって検討した」
- 「テキストを読んだ印象から判断した」
  - 「読んだ印象」が含意する判断の構造⇒暗黙知
  - 「モチーフ」を見出す理路⇒暗黙知
  - …
    - これまでの正当性判断はピアレビューに委ねられていた
      - 研究データ明示化の可能性／必要性は？
      - ⇒工学オントロジー等による記述の可能性
      - ⇒ディープラーニングによる分析との対比
- 「触ったら感触でわかった」

研究成果の直接的な基礎となったデータ

- 「ノートやメモをとって検討した」
- 「テキストを読んだ印象から判断した」
- 「触ったら感触でわかった」
  - 触感が含意する様々な情報⇒暗黙知
    - 「紙の質・材料」等
    - ⇒材料の違いから判断できる事柄
    - ⇒触感を構成する要素の探求と明示化
      - マイクロスコープ／材料・素材の分析
      - 点の分析と面での判断の差異⇒記述の粒度は？
      - 他のアプローチは？

「研究成果の直接的な基礎となったデータ」の構築と共有の可能性

- 「ノートやメモをとって検討した」
  - ⇒既存の人文学向けの国際的なアノテーション記述ルールが利用可能（TEIガイドライン、IIIFアノテーション等のデファクトの国際コミュニティ標準）
  - ⇒紙媒体でも稀に公開されたことがある（研究ノート等）
    - ⇒すでにデジタルでの公開例多数

- 「テキストを読んだ印象から判断した」
  - ⇒記述ルール自体はいくつかの手法が利用可能
  - ⇒作成・公開・共有する習慣の有無は？無い場合はどうするか？

- 「触ったら感触でわかった」
  - ⇒記述ルールが未整備
    - 手法・記述粒度・ターミノロジー…
  - ⇒公開・共有する習慣はないが、新アプローチであるため公開・共有の習慣を新たに作りやすいかもしれない
  - ※この種の暗黙知は他にも存在し得る

「研究成果の基礎となったデータ」の媒体

- Journal of Open Humanities Data
  - https://openhumanitiesdata.metajnl.com/
- Japanese Journal of Digital Humanities
  - https://www.jstage.jst.go.jp/browse/jadh/-char/ja
  - J-STAGE Data による
- Zenodo, GitHub repository…

- ⇒掲載可能な場所は徐々に広がっている。
  - しかし、まだ容易に掲載できるというわけにはいかない模様。

データ構築・共有に関する日本の課題

- （反省も込めて）日本での人文学側からのアプローチの弱さ
  - しかし最近では…
  - Unicode等の環境整備が進み国際的なアプローチが比較的容易になった
  - 国際標準へのアプローチも徐々に進んでいる
  - 関連コミュニティは徐々に広がりつつある
  - テキストデータの共有に関する国際ガイドライン(https://tei-c.org/)に関する入門書が刊行(2022)

- 日本での「研究図書館」の担い手
  - 欧米の有力（大学）図書館（＝研究図書館）では研究資料の高度化に研究図書館が重要な役割を担っている
  - ⇒「サブジェクト」への特化、「デジタル」の内製化の度合い、関連技術者コミュニティの形成
  - 日本ではどこがこれを担うのか？

- 政策的な支援の遅れ
  - NEHのODH（米国）やNFDI（ドイツ）、ERICのDARIAH、CLARIN、CLARIAH等（EU）等に比較すると横断的な支援が弱い
    - 個別にはいくつかの取組みがある／「研究」として科研費で推進される例は多いが持続性に課題
    - 一部の分野に関しては支援が開始された
      - 日本学術振興会人文学・社会科学ﾃﾞｰﾀｲﾝﾌﾗｽﾄﾗｸﾁｬｰ構築推進事業

Generative AI language tools ?

- 出力結果への評価と入力データの整備・制御
  - 「公式な」利用は不可能でも非公式には様々に利用される可能性
    - これまでにも誤情報は多く存在したが…
  - よりまともなデータを読み込ませることの有用性
    - 内容的な信頼性
    - 機械可読性の高度化（テキストデータ化～テキスト構造化）
      - 機械可読性自体の可塑性・変更可能性（例：異体字が使えるようになった）
    - ⇒人文学データの有用性

研究データを対象とする持続可能な研究環境へ

- データの持続可能性
- ツールの持続可能性
- 手法の持続可能性
- モデルの持続可能性
- …

持続可能な研究環境へ向けた課題

- データの持続可能性
  - 課題：データの利用条件、データ形式の共通化、データ形式の利用条件、データの精度や信頼性、データの性質や目的…
- ツールの持続可能性
- 手法の持続可能性
- モデルの持続可能性
- …

持続可能な研究環境へ向けた課題

- データの持続可能性
- ツールの持続可能性
  - 課題：ツールの利用条件、ソースコードの利用条件、フレームワークの選択、開発の継続性…
- 手法の持続可能性
- モデルの持続可能性
- …

持続可能な研究環境へ向けた課題

- データの持続可能性
- ツールの持続可能性

- 手法の持続可能性
- モデルの持続可能性
  - 課題：汎用性の獲得、個別ドメインの意義を高めるための個別性の獲得、

- …

持続可能な研究環境への解決策

- データの持続可能性
  - オープンな利用条件（あるいはFAIR原則）のデータ構築
  - 分野特化型と汎用性を兼ね備えた共通データ形式の開発と採用
  - データの精度や信頼性を確保するための記述ルールの開発と徹底
  - データの性質や目的の記述ルールの開発と徹底
  - …

- ツールの持続可能性
- 手法の持続可能性
- モデルの持続可能性
- …

持続可能な研究環境への解決策

- データの持続可能性

- ツールの持続可能性
  - 標準化されたデータ形式をターゲットとした開発
  - Research Software Alliance のコンテクスト
    - https://www.researchsoft.org/funders-forum/
  - 分野特化型と汎用性を兼ね備えた開発とモジュール化
  - …

- 手法の持続可能性
- モデルの持続可能性
- …

持続可能な研究環境への解決策

- データの持続可能性
- ツールの持続可能性

- 手法の持続可能性
- モデルの持続可能性
  - データやツールのオープン化・標準化との相互連携
  - 議論と合意形成のためのコミュニティ
- …

持続可能なデジタル研究環境に向けて

- データ・ツールのオープン化と標準化
- 汎用化とドメイン特化との相関的開発
- データやツールと手法・モデルとの相互連携
- 議論と合意形成のためのコミュニティ

公共的利活用に向けた“より適切な"データの作成

- 人文知の国際的な共有に向けたデータの構築
  - 国際的な枠組みに準拠したデータ作成と**枠組み自体の検討・改良**
    - 日本ローカルの国際標準への反映が国際標準における多様性を担保する
    - 国際的な枠組みへの準拠は独自ルールの維持運用コストを低減する
      - ⇒枠組みの構築・改良に向けた国際的な活動への積極的な参画の重要性
    - 対応ツールの開発が国際的な貢献にもつながる

- 人文知を適切に反映したデータの作成と共有知のベースとしての"Generative AI language tools"への貢献の可能性

  - 制御可能な"Generative AI language tools"は運用可能か？