

人文情報学とは何か

大 向 一 輝

1. はじめに

近年、人文情報学あるいは Digital Humanities (DH) と呼ばれる研究領域が注目を集めており、国内外で研究会や学術会議が開催されている。人文情報学はその名の通り人文学と情報学との学際領域だが、その内実については研究コミュニティとして統一的な見解があるわけではない。

人文情報学の代表的な拠点のひとつであるカリフォルニア大学ロサンゼルス校 (UCLA) の講義資料の冒頭には、“Digital humanities is work at the intersection of digital technology and humanities disciplines.” と記載されている⁽¹⁾。また、スタンフォード大学 Humanities Center のウェブサイトには、“The digital humanities at Stanford sit at the crossroads of computer science and the humanities.” とある⁽²⁾。人文情報学の入門書「Digital_Humanities」では、“Digital Humanities refers to new modes of scholarship and institutional units for collaborative, transdisciplinary, and computationally engaged research, teaching, and publication.” とあり、研究分野としての位置づけに留まらない、教育や出版の形態を含めた定義がなされている⁽³⁾。国内では、情報・システム研究機構人文学オープンデータ共同利用センターの北本氏のウェブサイトにおいて「デジタル・ヒューマニティーズとは、人文学的問題を情報学的手法を用いて解くことにより新しい知識や視点を得ることや、人文学的問題を契機として新たな情報学の分野を切りひらくことなどを目指す、情報学と人文学の融合分野である」との記述がある⁽⁴⁾。

いずれの記載でも、人文情報学は人文学研究におけるコンピュータやインターネットの活用といった一方通行の関係ではなく、双方向的あるいは複合的な状態として定義されていることが興味深い。一般に、多くの学問分野にとって情報学が取り扱う技術は単なる研究の一手段として捉えられているものと思われる。人文学においては、昨今の史資料のデジタル化によって研究者が扱うことのできる情報量は飛躍的に増大したが、基本的な研究手法は従来と大きく変わるものではない。情報学の側でも、その成果を利用する各専門分野固有の課題には深く踏み込まず、誰もが等しく恩恵を受けられるように汎用化された技術を高く評価する傾向がある。その帰結として、今日まで人文学と情報学は相互に独立した関係性が保たれてきた。人文情報学はそのようなあり方とは異なるものとして定義されているが、実際の研究ではこういった連携が行われているのか、以下に典型的な事例を示しながら議論を進めたい。

2. 人文情報学の事例

2-1. Google Books Ngram Viewer

Google は2003年より米国の大学図書館を中心とした蔵書のデジタル化を進め、書籍の全文検索・閲覧サービスである Google Books を公開している。Google Books は書籍の画像だけでなく文字情報のテキスト化を行っているが、2010年にはこれらのテキスト情報を統計的に分析し、時系列で可視化することができる Ngram Viewer (<https://books.google.com/ngrams>) を研究者向けに提供した。Ngram とは連続する単語、すなわち2単語であれば2-gram、3単語であれば3-gram を総称した概念である。Ngram Viewer を用いることで、特定の単語やその組み合わせとしてのフレーズがどの年代の書籍に何回出現するかを知ることができる。例えば「The United States are/have (複数形)」と「The United States is/has (単数形)」を比較すると、その頻度が1880年代から1900年代の間に逆転しており、米国に対する認識が連合国から単一国家へと変化したタイミングがこの時期であると推測することが可能である(図1)。このように、単一ないし少数の書籍を精読(close reading)するのではなく、同時代の、あるいは時代を問わず大量の書籍の中から機械的な処理によって知見を得る方法は「遠読(distant reading)」と呼ばれている⁽⁵⁾。遠読はコンピュータやデータベースが存在しなければ実現することが不可能な人文情報学の代表例として挙げられることが多い。

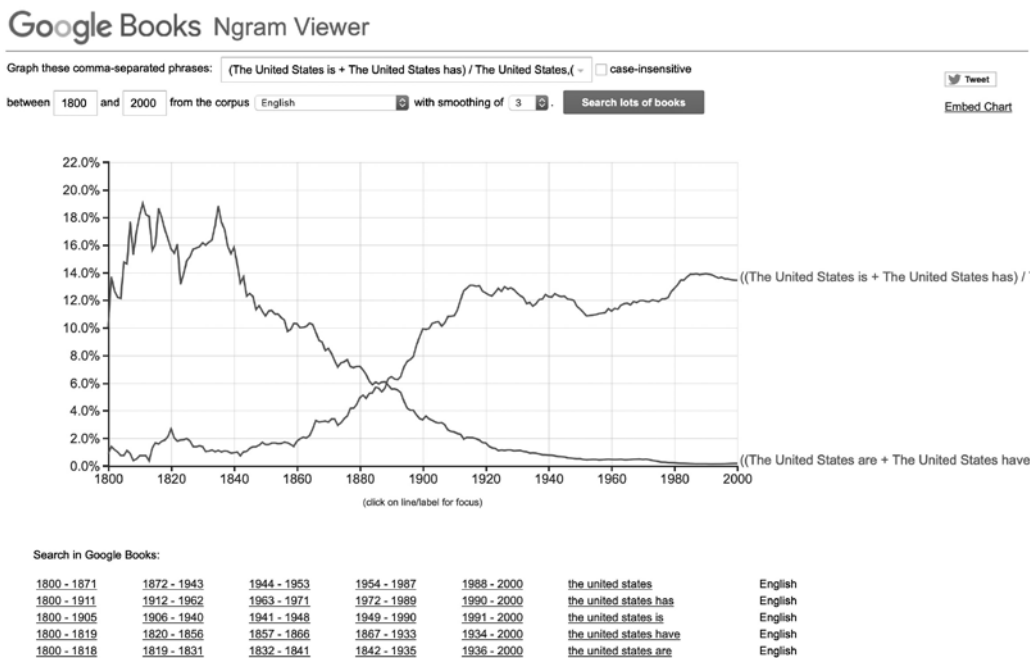


図1 Google Books Ngram Viewer による比較

2-2. 機械学習によるテキストマイニング

一群のテキスト情報の中から特定の表現の出現回数を統計的に把握し、複数の作品が同一著者によるものかを判定したり、著者の作風の変化を時系列的に分析するなどのテキストマイニングは人文学分野においても古くから実施されてきた。前述の Google Books Ngram Viewer もこのような研究を通時的かつ大規模に展開したものと位置づけることができる。一方、情報学分野では機械学習を用いた文書群の分類手法が提案され、迷惑メールの自動分類などで実用化されている。現在では、これらの手法を人文学資料に適用する研究が多数進められている。本稿では代表的な手法として確率的トピックモデルを取り上げる。確率的トピックモデルでは、1つの文書が複数の潜在的な話題（トピック）を含んでいること、またトピックの数が十分に用意されれば、文書群の中のすべての文書の内容はトピックの組み合わせで表現できることを想定する。その上で、個々の潜在的トピックを実際の文書中の単語から自動的に算出し、得られたトピックに基づいて文書の性質を明らかにする。この手法を南北朝時代の史料群に対して適用した研究では、30種類のトピックと、それぞれのトピックに含まれる単語が図2のように得られる⁽⁶⁾。トピックの中には、武士に関するもの（V1）、仏教関係者にまつわるもの（V9）などが散見される。同じようなトピックを含む文書同士は関連があると推測されることから、大量の文書群を再整理することも可能である。なお、この手法はトピックの総数を分析者が決めてから実施する。一般に、文書群に対して十分なトピック数は事前には予測できないことから、適切な結果を得るためには試行錯誤を行う必要がある。その後の研究では文書群の性質に応じてトピック数を自動決定するアルゴリズムも提案されており、課題の発生とその解決を繰り返す工学的な進展を見せている典型例でもある。

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
1 鎌倉之	天津	入道々	御房事	コノ一行	一々	一々	護言上	阿闍梨	阿闍梨	住人百	下村事	のあ	親王了	一々
2 神主十	入寺供	合戦之	執金	コノ一行	一々	一々	護言上	阿闍梨	阿闍梨	住人百	下村事	のあ	親王了	一々
3 足利將軍家	阿闍梨	御房日々	如件也	文書トシテ	本所之	同々	之由事	法印也	依会	有職免一	行清寄附	なり合	別当不知	已上一
4 足利將軍家	阿闍梨	御房日々	如件也	文書トシテ	本所之	同々	之由事	法印也	依会	有職免一	行清寄附	なり合	別当不知	已上一
5 第二	年預之	元弘三年七	護下	五郎代	前机	十二月中	可一々	護言上	阿闍梨	住人百	下村事	のあ	親王了	一々
6 範満公	入か	御奉書	候了	アリシ	同前	一々	候あ	護言上	阿闍梨	住人百	下村事	のあ	親王了	一々
7 九号文	寺々	進上之	僧正之	ナリト	吉村	三々	之聞	護言上	阿闍梨	住人百	下村事	のあ	親王了	一々
8 尊氏氏	一自	状如此候	如此之	コノ文書ノ	二〇	五十八	御々	護言上	阿闍梨	住人百	下村事	のあ	親王了	一々
9 夢重陸文	集會事	如件也	天氣	第一三	可為三分之	百人	可被下	護言上	阿闍梨	住人百	下村事	のあ	親王了	一々
10 文敏	行事也	同々	之由事	ホア	新条	五十枚初後之	如二	護言上	阿闍梨	住人百	下村事	のあ	親王了	一々
V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30
1 一反三	和尚也	状如此候	諸国々	優得	一々	日本与本省	一宇寺	御殿	一々	ヨリテ	寺之	敬白奉	次々	時真之
2 百歩ハ	答日道	寄進セラレ	当時ノモノ	授テ	不可也	本所	次二十	御殿	之聞	ヲモ	ケねん	陀羅尼田	朝臣也	正吉令
3 反小五	神師也	右件々	朝臣也	聽聞等	状如此候	日ク	まき	神田事	一通不可為ヘタ	一帳	真言之	云々者	正弘一期	之
4 科田三	進人	者也了	不可一紙	宮主之	年貢之	不該行	祝言過テ	御イセ	於ハ	シテ書	權別当法	光明寺	今日之	時友之
5 十歩下田二	云々	四至内	依進之	一職	請文之	建武三	續宣之	十七	為一	二あて	和印和尚	本尊之	晴儀	時友之
6 反六十	不一	限永代所	衆々	元年	百姓等不引	沛艾	序屋二	上分三	之由事	キリ	若宮之	然則下	中納言兼	時沢名文
7 二反三	如何候	相副之	不論国	六支一	時者也	實無	供を	六也	上者也	ヲハ王	都蘭	地蔵堂	大納言家	時沢名文
8 大々	為一	仍為後日	遠近之	劇説	可致之	於時真者	社務之	二丁未年	沙汰之	トラス	二内	然際於	阿弥陀三	之聞
9 小々	諸仏事	奉寄進之	以後之	又寺	請申也	天皇の	次日	外宮上	之果	ナカノ	黒田	阿弥陀三	之聞	以師兼敬
10 島か	以て	田地一反	永代所	衆々	元年	百姓等不引	沛艾	序屋二	上分三	之由事	キリ	若宮之	然則下	中納言兼

図2 南北朝史料における潜在的トピック

2-3. 時間基盤情報

史資料の内容の中には出来事の時間に関する表現が含まれていることがあり、複数の史資料の情報を相互比較することで前後関係の判別や新たな知見を得るための手がかりとなる。しかしながら、個々の時間表現の背後にある暦の体系が異なっている場合には単純に配列することができない。そこで、複数の暦を関連づけ、データベースとして提供する研究が行われている⁽⁷⁾。本研究では統一的な暦の表現手法としてユリウス通日を採用し、これとユリウス暦やグレゴリオ暦などの西洋の暦だけでなく和暦やイスラム暦などを対応させることで、多様な時間表現の相互変換を可能にしている（図3）。また、「鎌倉時代後期」や「2010年代」など、期間を有し、かつ曖昧な時間表現に関して明示的な記述規則を定義することで、複数の時間表現を論理的に正しく重ね合わせるなどの演算手法を提案している。

Time Information System **HuTime**

解析ソフトウェア | 時間基盤情報 | 資料・参考文献 | リンク | このサイトについて | English

ホーム > 時間基盤情報 > 暦変換 > 暦変換サービス

暦変換

和暦 | 暦 | ユリウス/グレゴリオ暦・1752年改暦（西暦） | 暦について

書式指定の前に型を選択 | 自動 | 書式指定と型 | 書式について

変換

暦日（始点）

左側のボックスに 平成25年4月1日 など、日付を表すテキストをそのまま入力し、変換ボタンを押してください。
複数の日付を入力する場合は、改行して次の行に日付を書いてください。100行までは入力できます。 → 使い方見本

Copyright © 2012-2019 Tatsuki Sekino. All rights reserved.

W3C XHTML+RDFa W3C CSS

図3 時間基盤情報の暦変換フォーム (<http://www.hutime.jp/basicdata/calendar/form.html>)

2-4. CiNii Books とデジタルアーカイブの連携

筆者がこれまで開発や運用に関わってきた学術情報サービスにおいても人文情報学の成果が生かされている。CiNii Books (<https://ci.nii.ac.jp/books/>) は国立情報学研究所が提供する、国内の大学図書館の総合目録を検索するための一般向けサービスである。従来、CiNii Books が検索結果として提示してきたのは書誌情報と各々の図書館における所蔵状況に限られていた。近年、国内外を問わず書籍や史料がデジタルアーカイブとして公開される事例が増加したことで、検索後にそのデジタル版を直接閲覧できる可能性が高まっている。このような連携を実現するためには、CiNii Books が持っている文献のメタデータと、各種デジタルアーカイブのメタデータが一致している必要がある。しかしながら、異なる機関で作成されたデータは採

録のルールや形式が異なっていることが多く、一致することは極めて稀である。実際には、様々なデジタルアーカイブにおけるメタデータ形式や登録ルールを確認し、同等の情報が記載される項目を洗い出した上で、機械的な変換によって情報の同一化ができるかどうか、データ登録の業務フローを見直すことが可能かなど、多面的な検討を行う。また根本的な解決手段としては、対象となる情報のすべてにグローバルかつ一意な識別子（ID）を付与し、このIDを用いて相互リンクを行うことが求められている。これを実現するにはIDを維持管理する情報システムとの接続や、業務フローの改訂を要するため、一部のデジタルアーカイブにおいて実証が進められている。

2-5. Linked Open Data による知識表現

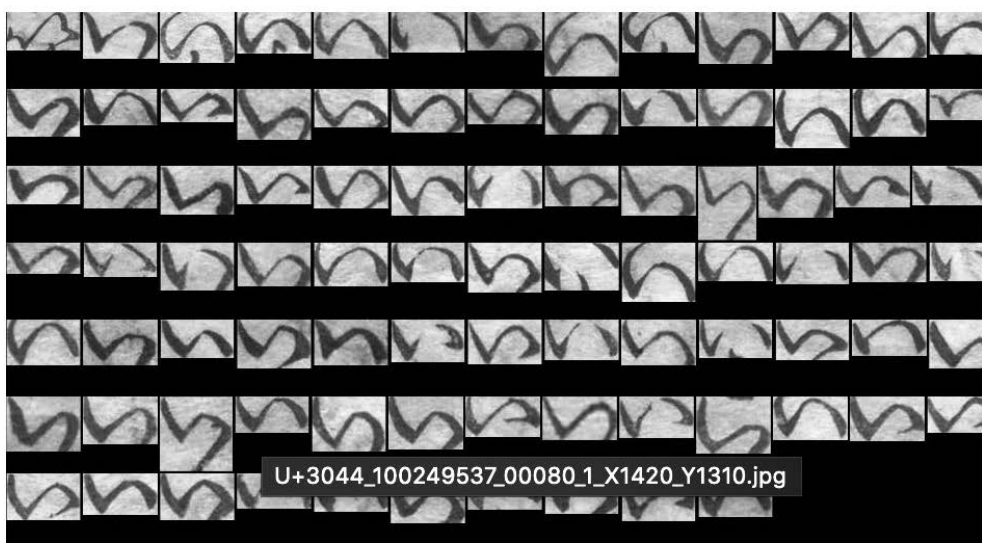
書籍や学術論文のように過去に整備された大量の情報が存在する場合には、前節の漸進的なアプローチによって徐々に改善を進めるべきである。一方、新たな分野の情報を作成する場合には、当初からあらゆる情報にIDを与え、ID同士の関係性を記述することによって知識や概念をネットワーク状に明示化する方式を選択できる。Linked Open Data（LOD）は、このような記述や公開のための方法論の総称であり、元来はインターネット上の知識表現や情報の自動処理を目的として提案されてきた。現在では、表現の自由度の高さや、異なる領域の知識であっても同様の形式で記述し、リンク付けが行える点が評価され、とくに欧米の人文学において知識表現ならびに提供手段として普及が進んでいる。代表例としては、Getty 財団による芸術・建築分野のシソーラス Art & Architecture Thesaurus（AAT）（<https://www.getty.edu/research/tools/vocabularies/aat/index.html>）や、国際的な図書館ネットワークで整備された人名典拠 Virtual International Authority File（VIAF）（<http://viaf.org>）が挙げられる。

2-6. 日本古典籍くずし字データセット

国内において古典籍の大規模デジタル画像化のプロジェクトが進められている。これに伴って、文字情報のデジタルテキスト化に対する要求が高まっているが、くずし字で書かれた文書は専門家による翻刻に頼らざるを得ず、その対象は著名な史資料に留まっている。情報学分野では人工知能技術の一つである深層学習の性能が飛躍的に向上したことで、画像からの物体認識や文字の読み取りが実用化されつつある。この状況を活用するために、人文学オープンデータ共同利用センターでは、くずし字画像を一文字ごとに切り出し、それらの画像に正しい読み情報を付加した「日本古典籍くずし字データセット」（<http://codh.rois.ac.jp/char-shape/>）を公開している（図4）。深層学習を含む機械学習では、大量の事例データ（ここでは画像）と正解データ（読み情報）を学習アルゴリズムに与えると、事例と正解とを関連づける数学モデルが自動的に生成される。用意するデータの量が多ければ多いほど精度が高まる。アルゴリズムの研究開発にはデータセットが必須であるため、質の高い大規模データセットを作成することで国際的かつ先端的な研究者を巻き込むことにもつながる。日本古典籍くずし字データセットは、世界的に利用されている手書き数字データセットと同じ仕様で作成することで、言語圏

を問わない研究コミュニティの参加を目指している。

料理珍味集 (89)



好色一代男 (746)

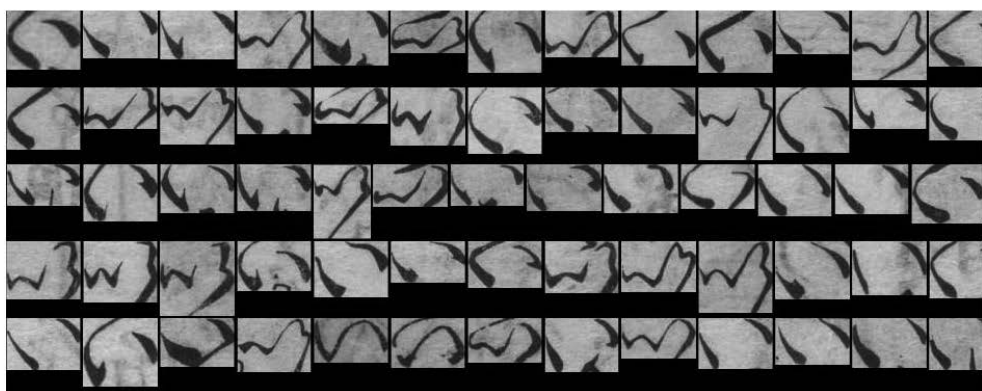


図4 日本古典籍くずし字データセット (一部)

2-7. クラウド翻刻

コンピュータによる自動的な翻刻とは異なるアプローチとして、専門家以外のコミュニティにテキスト化の作業を委ねる、いわゆるクラウドソーシング翻刻も普及しつつある。国内のクラウド翻刻の成功例としては、京都大学古地震研究会による「みんなで翻刻」(<https://honkoku.org>)がある。みんなで翻刻では、地震に関連する史料の翻刻を一般ユーザが行えるようにインターネット上のサービスを提供し、300人以上の協力者を得て約500点の史料の翻刻を2年間で完了した実績がある。翻刻した文字数を参加者間で競うランキング機能をはじめとしたコミュニティ活性化のための機能を備えるだけでなく、初学者がくずし字の読み方を学ぶ

ためのワークショップを継続的に開催するなど、幅広い活動を行うことで裾野を広げ、目標を達成していることが特徴的である。海外でも米国議会図書館による「By the People」(<https://crowd.loc.gov>) のように図書館が主導するもの、「Transcribe Bentham」(<https://blogs.ucl.ac.uk/transcribe-bentham/>) のように特定の人物の著書や書簡に特化した研究プロジェクトなど、クラウド翻刻は多様な展開を見せている。

3. 課題と展望

ここまで、人文情報学を象徴するような研究や活動を紹介してきた。ごく少数の事例ではあるが、人文情報学が持つ特性として、1) データ量が大规模であること、2) 多層的な情報を扱っていること、3) 共有と協働が志向されていることがわかる。これらの特性は、デジタル化によってコンピュータ上で情報を扱うことが可能になり、自由にコピーができるようになったこと、さらにネットワークを通じて他のコンピュータとその利用者に対して容易に情報を送信できることではじめて実現される。この点において、情報学あるいは情報技術は専門分野の単なる道具としての位置づけを超えた役割を担っていると言える。

一方で、人文情報学の特性や方法論はさまざまな面で批判を受けている。例えば、2-1節で紹介した Google Books Ngram Viewer で明らかになった何らかの傾向（19世紀終盤に連合国から単一国家へと米国の認識が変化したことなど）について、その理由をこのサービスから直接知ることはできない。この傾向は意味が捨象された一種のパターンであり、解釈可能性・説明可能性が欠如している。大量のデータを機械的に処理することと、その内実を理解するための努力は互いに関係しておらず、人文情報学の成果が従来の人文学を上書きするような状態にはなっていない。あくまでも大規模データ処理は人文学研究者にとってインスピレーションを得るための手段でしかなく、道具としてのコンピュータの延長線上に過ぎないという見方がある。

あるいは、2-3節の時間基盤情報のように、史資料間を横断して統制すべき概念は時間の他にも地名や人名、組織名など多数に上る。これらすべてを明示化し、あらゆる史資料に適用することができれば情報量が飛躍的に増大するが、このような地道かつ膨大な作業に専門家自らが従事することは現実的でない。研究に対する貢献の大きさが予測できない状況において構造化データの作成に邁進することは費用対効果の面でも得策ではない。

2-7節のクラウド翻刻については、得られた結果の信頼性について疑念が呈されることが多い。協力者の数が増えたとしても、翻刻結果に誤りが含まれるのであれば学術的には価値がない。また多くの人々を巻き込むための労力が極めて大きいため、常に専門家がプロジェクトに主体的に関わることができるかどうかは定かではない。

どの批判も、単体としての研究活動に対しては的を射ており、より高度な情報技術を導入したとしても改善するものではない。人文情報学としては、個々の批判に対して、複数の特性を組み合わせることによって解決する道を探るべきであろう。大規模データ処理については Ngram のように単語の連なりだけを入力するのではなく、これまでの研究で積み上げられて

きた構造化情報を与えることによって意味付けや解釈を可能にする手法が求められる。また、多層的な情報の明示化作業にはクラウドソーシングが適しているものと思われる。クラウドソーシングの信頼性確保のためには、複数の作業者による結果の相互比較や、過去の作業実績データから信頼性の高いユーザを把握し、確信度を上げるといった大規模データの活用が有効であろう。

このように、情報技術をもっぱら個別の課題解決に用いるのではなく、相補性を考慮した全体のシステムの中に位置づけることが必要である。今後の人文情報学が意義ある成果を挙げるためには、設計者たる人材の育成も重要な課題となるだろう。

4. 人文情報学の教育プログラム

最後に、人文情報学が学問分野として成立し、多様な人材を輩出するために必要不可欠な教育プログラムについて述べる。現時点において人文情報学は分野としての日が浅く、全世界的に共有された教育プログラムは存在していない。本稿の冒頭で取り上げた UCLA の講義資料では、学ぶべき基礎技術として展示（デジタルアーカイブ）、データ管理、可視化、テキスト分析、地図・時間情報、インターフェイス設計、HTML が取り上げられている。また、サセックス大学の Berry 氏は人文情報学の構成要素を階層的に表現した「Digital Humanities Stack」を提示している⁽⁸⁾(図5)。Digital Humanities Stack は教育プログラム自体ではなく、コンピュータ的思考や知識表現といった基本概念からアーカイブ、メタデータ、そしてツール、アプリケーション、出版に至るまでの多様な要素を整理している。

人文情報学では従来の人文学が扱う専門知識に加えてこれらの概念や技術を習得する必要がある、その分量は膨大なものとなる。現実的には研究活動を行うために必要最小限の技術のみを学ぶことになると思われるが、それらの技術がどのようなコンテキストで生み出されたのかといった背景知識やコンピュータ的思考の概要については基礎的なカリキュラムの中に組み込む必要がある。その上で、人文学研究での活用あるいは人文情報学ならではの研究へと発展させていくためには、研究対象や手法の違いに関わらず情報技術について互いに教え合い、切磋琢磨できるコミュニティや場の存在が重要になるだろう。

The Digital Humanities Stack

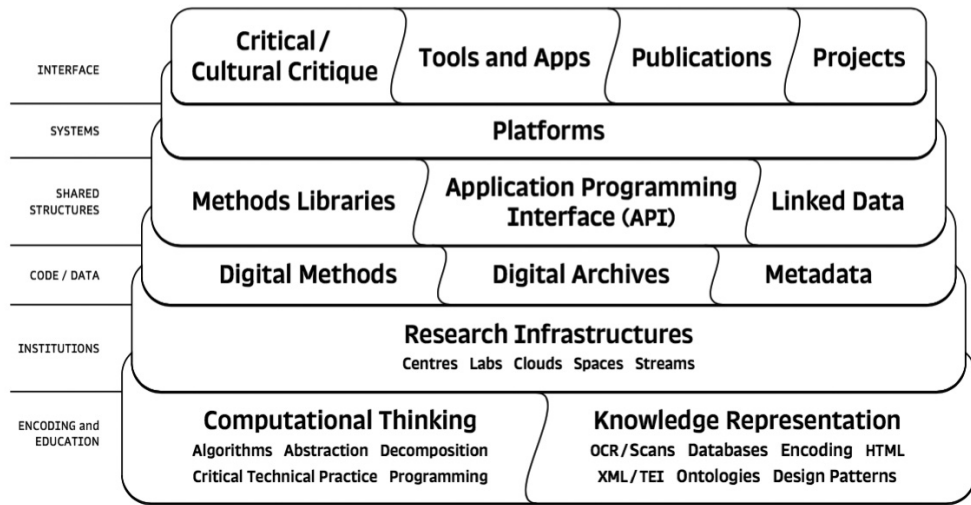


図 5 Digital Humanities Stack

参考文献

- (1) J. Drucker: Introduction to Digital Humanities - Concepts, Methods, and Tutorials for Students and Instructors (2014).
http://dh101.humanities.ucla.edu/wp-content/uploads/2014/09/IntroductionToDigitalHumanities_Textbook.pdf (accessed 2020-01-06)
- (2) <https://shc.stanford.edu/digital-humanities> (accessed 2020-01-06)
- (3) A. Burdick, J. Drucker, P. Lunenfeld, T. Presner and J. Schnapp: Digital_Humanities. MIT Press (2012).
- (4) <http://agora.ex.nii.ac.jp/~kitamoto/research/dh/> (accessed 2020-01-06)
- (5) F. Moretti: Distant Reading. Verso (2013).
- (6) 山田太造, 野村朋弘, 井上聡: 日本南北朝期史料を対象とした潜在的トピックによる史料分類と関連史料提示の手法. 情報処理学会シンポジウムシリーズ, Vol. 2013, No. 4, pp. 145-152 (2013).
- (7) 関野樹, 山田太造: 日付を表す文字列の解釈と暦の変換—暦に関する統合基盤の構築に向けて. 情報処理学会シンポジウムシリーズ, Vol. 2013, No. 4, pp. 161-166 (2013).
- (8) D. Berry and A. Fagerjord: Digital Humanities: Knowledge and Critique in a Digital Age. Wiley (2017).

