歴史研究者のためのデジタル道具箱

小風 尚樹

はじめに

最近、囲碁将棋や車の自動運転などで注目を集めている人工知能 (Artificial Intelligence, AI) の分野で、「シンギュラリティ（Singularity: 技術的特異点）」という言葉がある[1]。定義は諸説あるが、一般に流布する意味内容としては、さながら SF フィクションのような、人工知能が賢くなって人間の知能を遥かに上回る時点のこととされる[2]。

このシンギュラリティの議論と関連して話題になったのは、オクスフォード大学のオズボーン (Michael Osborne) 氏による「雇用の未来――コンピュータ化によって仕事は失われるのか」という論文で、702 の職種の特性調査に基づいて、近い将来なくなる確率の高い職業がランキング形式で発表されたことである。金融関連業務やデータ・機械の取り扱い、あるいはサービス業に至るまで様々な職業が挙げられており、オズボーン氏によれば、人間が失業しないためには高い社会的技能と創造性が必要であるという[3]。このように様々な領域の職業において、テクノロジーの進歩に伴い、人間とコンピュータのタスク分化が進んでいくことが指摘されているが、果たして社会における人間とコンピュータの適切な向き合い方とはどのようなものになっていくだろうか。

こうした AI ブームの時代において、人間による人文学研究がどのように特徴づけられるのか、いま一度考え直してみたい[4]。

そこで本稿では、特に歴史研究に関する作業そのものに着目し、コンピュータが研究者のタスクをどの程度代替できるのか、あるいはそれらに関連する技術が今後発展し得る方向性について検討する。こうした情報通信技術の進歩は、人間が行う人文学研究の作業の見直しを迫る外的圧力として働く一方で、これまでの人文知の蓄積をよりよく表現できる可能性を有している[5]。本稿が、研究を取り巻く技術的環境の現状について情報

---

[1] Ray Kurzweil, *The Singularity is Near: When Humans Transcend Biology*, Viking Press: New York, 2005（井上健監訳『ポスト・ヒューマン誕生――コンピューターが人類の知性を超えるとき』NHK 出版、2007 年；NHK 出版編『〈エッセンス版〉シンギュラリティは近い――人類が生命を超越するとき』NHK 出版、2016 年）.

[2] 塚本昌彦「ウェアラブルからシンギュラリティへ」『情報処理（情報処理学会誌）』第 57 巻第 10 号、968-969 頁。

[3] Carl Benedikt Frey and Michael A. Osborne, ‘The Future of Employment: How Susceptible are Jobs to Computerisation?’ http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf (accessed 1st April 2017), September 2013, pp. 1-72. 以下、URL のアクセス日時は同じであるため、省略する。

[4] ここで、広く人文学研究と銘打っているのは、2016 年の日本西洋史学会でもシンポジウムが開催されたデジタル・ヒューマニティーズの潮流を意識してのことである。この学問分野の情報源などについて紹介した論考として、以下を参照のこと。菊池信彦「デジタルヒューマニティーズ／デジタルヒストリーの情報源――デジタル時代の歴史学を考えるために」『現代史研究』第 59 号、2013 年 12 月、55-68 頁。

[5] 人文学研究におけるコンピュータ技術の適用は、このような外的圧力と同時に、そもそも研究上の必要性という内発的な動機によっても進められてきた。Jennifer Edmond, ‘Collaboration and

を提示することによって、読者諸氏の研究に関する新たな気づきにつながる一助となることを願ってやまない。

各論では、基本的に高度なプログラミング技術を必要としないような web サービスやデジタルツールを紹介する。以下で取り上げる研究作業は、歴史研究者が史料にアクセスし、史料を読解し、研究成果を可視化するという過程に沿っている。具体的には、画像比較、手稿文書翻刻、統計的なテキスト分析、インタラクティブ[6]な地図作成を扱う。

画像比較

かつて西洋史研究者は、自らの研究対象地域に留学や史料調査に行く際に船で移動し、直筆の手紙による情報交換を行っていた。現在では、インターネットが普及し、文書館が所蔵する史料の問い合わせを迅速に行うことができるだけでなく、欧米であれば飛行機の直通便で半日もあれば赴くことができる。このような交通・情報インフラの整備によって、情報収集にかかる時間が大幅に短縮された環境に身を置く現在の研究者は、歴史研究における「新人類」とさえ称されることもある[7]。共同研究プロジェクトや国際学会の運営などについても、メールや SNS などのコミュニケーションツールの発展によって、仕事の幅は圧倒的に広がっている。情報インフラの恩恵を受けることなく研究活動に従事する歴史研究者はほぼ皆無と言ってよいだろう[8]。

さらに今では、自宅に居ながらにして、世界中の学術機関が公開する画像をパソコンの web ブラウザ上で同時に比較し、拡大縮小や回転を自由に行い、註釈を付与することもできてしまう。これには、デジタルアーカイブにおいて画像を公開し共有するための国際規格トリプル・アイエフ (IIIF: International Image Interoperability Framework[9]) 形式でデータ公開を行っている機関のデータと、画像閲覧のためのビューワ（フリーソフトウェアのものが複数公開されている）が必要である[10]。

例として、以下に 2016 年 5 月にオクスフォード大学ボドリアン図書館のデジタル・スカラシップセンターにて開催された IIIF ワークショップのイメージ画像を挙げる[11]。こ

Infrastructure', in Susan Screibman, Ray Siemens, and John Unsworth, eds., *A New Companion to Digital Humanities*, Wiley Blackwell: Malden, MA., 2016, pp. 54-65.

[6] インタラクティブな web ページとは、閲覧するユーザーのアクション（マウスによるクリックやスライド、あるいはタブレット画面のスワイプなどの操作）に応じて出力を返すようにプログラムされたものである。

[7] 深沢克己「『あまりに遠し』から西洋史研究の『新人類』まで（コラム 歴史の風）」『史学雑誌』第 124 編第 1 号、2015 年 1 月、40-42 頁。

[8] 長野壮一「〈デジタル時代の歴史学〉デジタル歴史学の最新動向――フランス語圏におけるアーカイブ構築およびコミュニティ形成の事例紹介」『現代史研究』第 61 号、2015 年 12 月、39 頁。

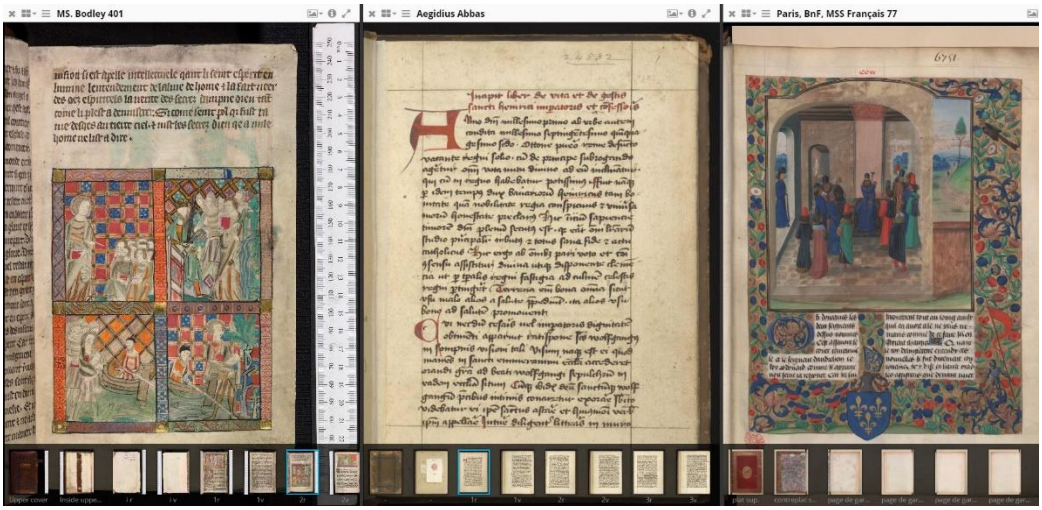
[9] http://iiif.io/

[10] mirador JS という IIIF 画像ビューワの使用法を解説した記事として、以下を参照のこと。オクスフォード大学ボドリアン図書館とデジタル・スカラシップセンター共催のワークショップ原稿 https://docs.google.com/document/d/1kFLuZV4t7IoMBqLNqiPpeJOtvJDeGxrg8fSIbAM3mYM/edit; 人文情報学研究所の永崎研宣氏によるブログ http://digitalnagasaki.hatenablog.com/entry/2016/04/28/192349

[11] Emma Stanford, 'Making the most of digitized manuscripts: a free IIIF workshop', https://www.slideshare.net/EmmaStanford1/making-the-most-of-digitized-manuscripts-iiif-and-the-digital-

のように、図像や文書史料など異なる種類の史料を比較検討できるだけでなく、表示画面も多く設定できるため、同じ史料群の写真を複数同時に閲覧することもできる。後者のような利用方法は、例えば絵巻物のように、もともと分割して見ることを想定していないような体裁の史料の分析などに役立つだろう。

図１　IIIF と mirador JS ビューワの活用イメージ画像[12]



図にもあるように、IIIF を採用している機関としては、La Bibliothèque nationale de France, British Library, British Museum, Cambridge University, DPLA, Europeana, Harvard University, Internet Archive, University of Oxford (Bodleian Library) など、西洋史研究者としても重要な顔ぶれが多く揃っている。参加機関は英語圏だけに限らないので、自身の研究に関連するデジタル画像を求めて、ぜひ脚註に記した URL から一覧をご確認いただきたい[13]。

手稿文書翻刻

日進月歩の画像認識技術は、我々の日常生活にも大きく浸透している。例えば、クレジットカードなどのデータ入力を人間が行っていることを証明するために、コンピュータが認識しにくい文字列の画像をユーザーに解読させて、クラウドソーシング[14]の発想で文字認識の精度を向上させるための reCAPTCHA システムには、すでに日常的に触れたことがある人も少なくないだろう。

---

manuscripts-toolkit, published 27th May 2016.

12 https://blogs.bodleian.ox.ac.uk/digital/tag/iiif/

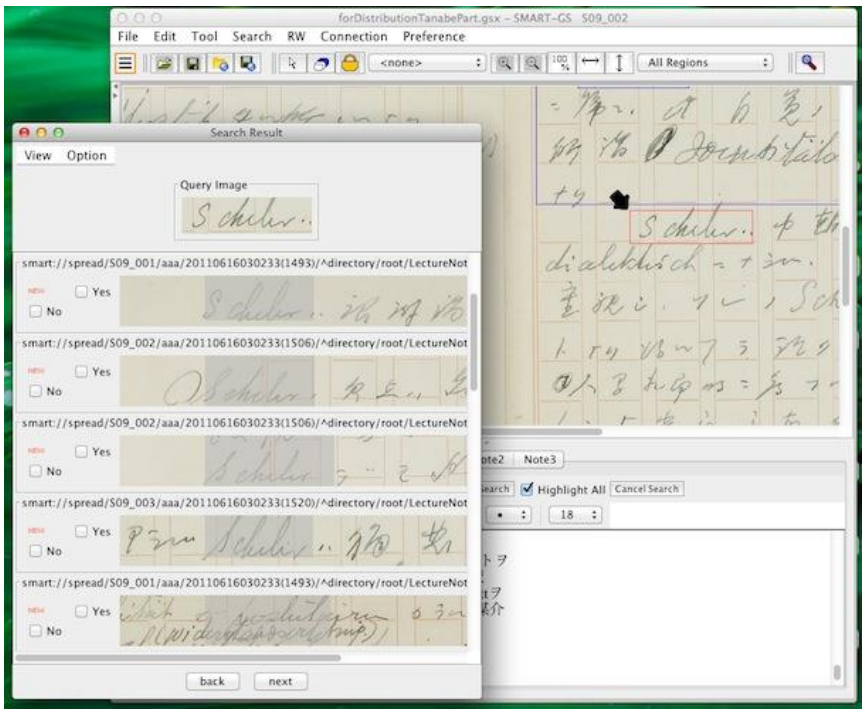
13 http://iiif.io/community/#participating-institutions

14 web 上で不特定多数の人々の協力を募り、サービスやアイディア、コンテンツを集約的に完成させることを目指す手法のこと。例としては、文化資源の面では、国立国会図書館デジタルコレクション（旧：近代デジタルライブラリー）でデジタル公開されている史資料の翻刻プロジェクト「翻デジ」が挙げられる。永崎研宣「日本語クラウドソーシング翻刻に向けて（特集 デジタル時代の日本語）」『情報の科学と技術』第 64 巻第 11 号、2014 年、475-480 頁。

図 2　reCAPTCHA システムの例

歴史研究の作業に関連してここで紹介するツールは、デジタル画像化した手稿文書を共同翻刻するツール、SMART-GS である[15]。このツールは、共同翻刻を支援するために、画像認識技術によって似た字形を検索する機能を提供している。OCR（Optical Character Recognition: 光学文字認識）技術のように、コンピュータが文字を解読してくれるわけではなく、あくまで翻刻支援のための技術であるが、似た字形が史料の中でどう出現するかを知ることができる。人名や地名の検索をもとに、索引を作成することも可能だろう。

図 3　SMART-GS の手書き文字サーチ機能[17]



しかし、手書き文字については、遠くない将来コンピュータが自動翻刻できるようになる可能性もある。このことに関連して留意しておきたいのは、Google の研究チームが、コンピュータに「猫」を高い精度で画像認識させることができるようになったというこ

[15] システムの狙いや使い方などの詳細については、開発者からの紹介に譲りたい。橋本雄太「集合知で読む歴史史料――SMART-GS が実現するグループリーディング」『人文情報学月報』第 37 号【前編】http://www.dhii.jp/DHM/DHM37_smartgs、2014 年 8 月。
[17] 同論文。

とである[18]。この研究成果の技術的な要点は、ディープラーニング（深層学習）という手法に基づき、動画共有サイト YouTube から取得した 1000 万枚の二次元の画像を学習材料としてコンピュータに読み込ませると、それらの画像から「猫」の特徴を自ら発見し（特徴量抽出）、「猫」の顔についての概念（ソシュールの記号論で言うところのシニフィエ）を獲得することに成功したという点にある。

図 4　ディープラーニングによって獲得された「猫」の顔イメージの可視化[19]

我々人間が「猫」を認識するようになる過程では、顔の特徴以外にも、体毛の手触りや色、鳴き声や匂い、一緒に遊んだ体験など、様々な情報を総合している。かたや AI は、二次元の画像情報だけを基に学習したため、サンプルは大量に必要となったのである。こうしたディープラーニングの技術は様々な分野で活用されており、最近では日本語古典籍におけるくずし字の解読にも応用されている[20]。字形の癖や連続して綴られる文字列の特徴を見つけ出すことは、ディープラーニングの得意分野であり、「猫」の複雑なシニフィエを獲得するよりも容易いと言われている[21]。

このように、手稿文書を解読する技術の基礎は徐々に出揃いつつあるように見える。いずれは、職人的技能を有する専門家以外に対してもこうした史料を利用することを容易にするかもしれない。しかしながら、技術的課題が克服されたとしても、サンプルとなる画像の提供と所有先の公開ライセンス、プロジェクトを進めるための資金等の様々な問題が立ちはだかることになる。これらについて本稿では立ち入らないでおこう[22]。

---

[18] Quoc V. Le, et al., ‘Building High-Level Features Using Scale Unsupervised Learning’, *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh: Scotland, UK, 2012, pp. 1-11, http://static.googleusercontent.com/media/research.google.com/en//archive/unsupervised_icml2012.pdf.

[19] Ibid., Figure 6. Visualization of the cat face neuron, p. 6.

[20] 拙稿「イベントレポート――じんもんこん 2016」『人文情報学月報』第 65 号【後編】http://www.dhii.jp/DHM/dhm65-2、2016 年 12 月。

[21] 松尾豊『人工知能は人間を超えるか――ディープラーニングの先にあるもの』KADOKAWA、2015 年、144-178 頁。同書は、これまでの AI ブームの歴史や今後の AI と人間の関わり方などについても、要点をまとめて解説している。

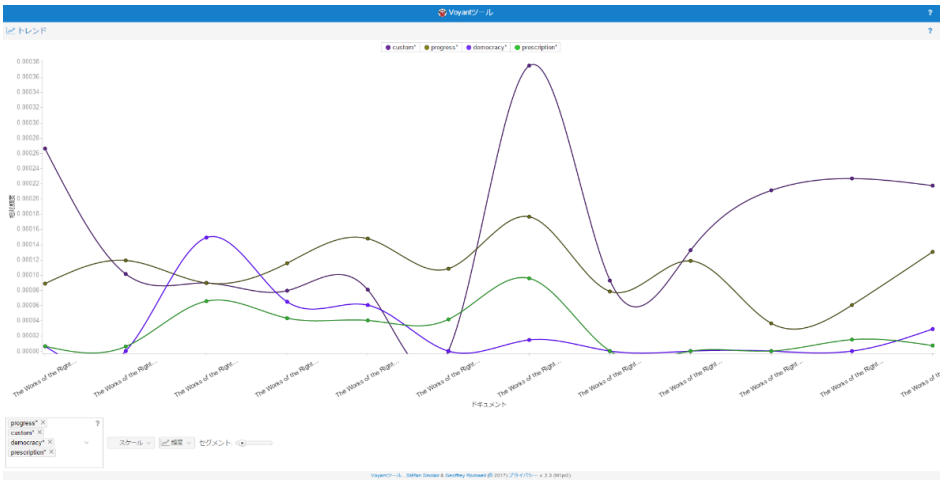
[22] 知的所有権などについては、以下を参照のこと。永崎研宣「人文学分野とサイバーインフラストラクチャ――デジタル・ヒューマニティーズにおける現状と課題」『情報の科学と技術』第 63 巻第 9 号、2013 年、369-376 頁。

統計的なテキスト解析

次に、テキスト解析のための Voyant-tools を紹介する[23]。これは、マギル大学のシンクレア (Stefan Sinclair) 教授とアルバータ大学のロックウェル (Geoffrey Rockwell) 教授による共同開発の成果であり、2016 年のアップデートでは、単語同士の間に余白がない日本語など世界中のあらゆる言語の解析にも対応するようになった。このアップデートを機に、Voyant-tools のメニュー画面が各国語に翻訳されたこともあり[24]、今後ますます利用が広がっていくと思われる。

筆者は以前、単一のテキストファイルを対象に、頻出語のリスト表示や、ある単語と共に用いられる共起語の分析、文脈の中での単語の出現分析（Key Word In Context 機能）といった、Voyant-tools が提供する基本的な機能の紹介を行ったが[25]、以下では、アップデートによって可能になった複数テキストファイルの統計的分析を行ってみたい。

図 5　Voyant-tools を用いた複数テキストファイルの word trend 分析



この図は、近世イギリスの政治思想家・哲学者・政治家として知られるエドマンド・バーク (Edmund Burke) の著作集全 12 巻のテキストデータ（全 12 巻の単語数の合計は、1,728,861 語）を対象に、バークの思想に関する語彙が、それぞれの著作集の中でどのような頻度で用いられているかを比較したものである[26]。ここでは、例として選んだ custom, progress, democracy, prescription の相対頻度が曲線で示され、横軸には巻数の若い順に左から著作集が並んでいる。このように膨大な量のテキストデータを手元に用意できるならば、精読を行う前の予備的な作業として、そのテキスト群の概略的情報を知ることが

---

[23] Voyant-tools, http://voyant-tools.org

[24] 日本語訳は、東京大学大学院の王一凡・佐藤正尚・杉浦清人・鈴木親彦諸氏と筆者、および永崎研宣氏により作成された。

[25] 拙稿「歴史研究とデジタル・ヒューマニティーズの相関」『東京大学学術機関リポジトリ』http://repository.dl.itc.u-tokyo.ac.jp/dspace/handle/2261/60849, 2016 年 9 月、1-8 頁。

[26] テキストデータは、Project Gutenberg (http://www.gutenberg.org/) から、Edmund Burke と検索し、*The Works of the Right Honourable Edmund Burke* の 1-12 巻の utf-8 形式のテキストファイルをダウンロードした後、本文に入る前の著者や目次情報をテキストエディタ（Windows の「メモ帳」など、テキスト編集のためのアプリケーションのこと）上で除去したものを用いた。

でき、先行研究で注目されてこなかった箇所が明らかになることもあるだろう。もちろん、あらかじめ得られた数量的な解析結果を鵜呑みにせず、質的な分析結果とも併せて考察する必要がある。

他にも､用例調査には非常に大きく貢献すると思われる｡たとえ同じ単語であっても、時代や地域、社会階層などによって込められた意味は異なるため、統計的なテキスト解析を行うことで、単語の分析やコロケーションの調査が進み、辞書の用例や意味が更新される可能性は大いにあり得る。そして最も重要な点は、こうした作業がもはや個人の手でできるようになってきていることである。

インタラクティブな地図作成

歴史研究の成果発信にあたっては、文章表現だけでなく、図表や地図を用いた表現が適切な場合も多くある。地図上での可視化を支援する web サービスは数多くあるが[27]、ここでは Neatline を紹介したい。Neatline は、学術機関が所有するコレクションを web 公開するためのプラットフォームである Omeka の拡張機能で[28]、時空間情報の表現に適したものである。

例えば、ホメロスの『イリアス』で言及された船舶の名称をギリシアの地理的情報に対応させたヴァージニア大学図書館によるプロジェクトのイメージ画像を下に挙げる。ここでは、本文中で船舶の名前が色付けされており、それらにカーソルを合わせると、地図上で対応する円が塗りつぶされる仕組みになっている[29]。

図 6　Neatline で作成された『イリアス』中で言及された船舶のマッピング

従来、多くの研究成果は紙媒体に載せて発表されてきたが、デジタルツールを用いればこのように多くの情報量を一枚の地図だけで表現できるようになっているため、研究成果を掲載する媒体が多様化する可能性についても念頭に置く必要があるだろう。

---

27 例えば以下のブログでは、データ可視化を目的とした 38 のデジタルツールが挙げられており、その中には、法人向けに日本語翻訳サービスが出された cartoDB などが紹介されている。
http://www.creativebloq.com/design-tools/data-visualization-712402

28 Omeka, http://omeka.org/

29 Jenny S. Clay, Courtney Evans, and Ben Jasnow, *Mapping the Catalogue of Ships*, http://ships.lib.virginia.edu/, Scholar's Lab at the University of Virginia Library.

結び

以上、紹介してきたように、人間では処理しきれない多くのデータをコンピュータに処理させることによって、一見すると明らかでない関係性が見えてくることもあるが、これはやはり「書かれているもの」を基礎としている点が重要である。一方で、人間の行う人文学研究の一つの大きな特徴は、「書かれていないものを読み解く」ことではないか。社会状況や文化規範といった、過去に生きる人々が持っていた暗黙知を再構成することが歴史研究の一つの重要な貢献だとすれば、人間による人文学研究は、やはりこうした点にこそ特徴がある。

本稿で紹介したようなデジタルツールを、研究に取り入れることに抵抗感を覚える向きもあるだろう。もちろん筆者も、例えば手稿文書を解読できるようにならなくて良いなどとは考えていない。高度な情報学的方法論を取り入れた人文学研究は、そもそも学際的な研究であり[30]、こうした成果に対する評価体制の整備は、国際的にみても端緒についたばかりである[31]。研究手法として採用するには、慎重を期す必要がある。

しかしながら、我々はすでにこの高度に情報化された社会に生き、あるいは巻き込まれており、情報への向き合い方を考えることを常に強いられている。歴史上の情報についても、技術的環境の変化から超然としているわけにはいかないだろう。今やチェスの世界では、人間とコンピュータがどのように手を組んでも良いフリースタイルの国際大会が開かれているように、社会における人間とコンピュータの付き合い方が変わっていくことが予想される。研究における作業についても、我々人間が創造的な知見を生み出すべく、コンピュータの計算能力を可能な限り活用する時代になってきているのではないだろうか。

本稿の執筆にあたり、国内外でデジタル・ヒューマニティーズの普及に 20 年近く努めておられる人文情報学研究所の永崎研宣先生からは、専門的な助言を賜った。原稿の企画・校正段階では、フランス留学中の長野壮一氏や山王綾乃氏も、デジタル・ヒューマニティーズに関心を寄せる若手西洋史研究者の立場からご意見をくださった。ここに記して感謝の意を表します。

---

[30] 学問の枠組みを担保するための制度として査読システムが挙げられるが、情報学的手法を取り入れた人文学研究に対しては、分析の妥当性を判断するための人員が恒常的に揃っているわけではない。Cf. Kathleen Fitzpatrick, ‘Peer Review’, in Susan Screibman, et al., eds., *op. cit*., pp. 439-448.

[31] 2015 年 9 月には、アメリカ歴史学協会によって、デジタル手法を用いた歴史研究を評価するためのガイドラインが発表された。https://www.historians.org/teaching-and-learning/digital-history-resources/evaluation-of-digital-scholarship-in-history/guidelines-for-the-professional-evaluation-of-digital-scholarship-by-historians. 同ガイドラインは、日本におけるデジタル・ヒューマニティーズの研究成果に対する学術的評価をめぐる議論を活性化させるために、日本語訳も行われた。菊池信彦・小風尚樹・師茂樹・後藤真・永崎研宣訳「歴史学におけるデジタル研究を評価するためのガイドライン」『東京大学学術機関リポジトリ』http://repository.dl.itc.u-tokyo.ac.jp/dspace/handle/2261/59142、2016 年 2 月、1-9 頁。