東アジア人文情報学の可能性についての試論 ： デジタルテキスト構造化の動向を中心として

東アジア人文情報学の可能性についての試論
――デジタルテキスト構造化の動向を中心として

永　崎　研　宣

Possibilities of East Asian Digital Humanities:
Through the Activities of Text Encoding

NAGASAKI Kiyonori

This paper discusses the possibility of DH in East Asia through a survey of past trends in text encoding in East Asia, particularly Japan. In Japan, the construction of text databases has been active since the 1980s, and a large number of text databases of Chinese classics have been created in China and Taiwan. However, the leading standard of text encoding, the TEI (Text Encoding Initiative) guidelines were not popular at that time. In the 2010s, as several major barriers have been reduced, the TEI guidelines have become more globally oriented, and a system for dealing with East Asian texts has been established. In the future, it will be possible to produce TEI-compliant structured texts in a more appropriate form, and they will be connected with multimedia data to promote various kinds of digital research. This will lead to a truly global digital research environment in the humanities.

キーワード：Unicode、TEI ガイドライン、IIIF、マークアップ

はじめに

DH（デジタル・ヒューマニティーズ、人文情報学）は、人文学においてデジタル技術を応用して研究を発展させるための一連の営みを指している。複数のオンラインジャーナルの運営に携わる国際 DH 学会連合（ADHO, Alliance of Digital Humanities Organizations）がオックスフォード大学出版局から刊行する基幹ジャーナルの名称が Digital Humanities ではなく Digital Scholarship in the Humanities、すなわち、人文学におけるデジタル学術研究、となっているのは、DH がどちらかと言えば人文学を志向するものとして推進されてきたことを端的に示していると言えるだろう。

DH という言葉が生まれたのは 2004 年頃のことである。この頃に刊行が企画された Humanities Computing の入門書に、出版社の編集者からの提案により A Companion to

Digital Humanities という名称が付けられたことがその起源である。この名称はその後、欧州や北米の学術政策で戦略的に使用され、国際 DH 学会連合の立ち上げや政府レベルでの研究資金や研究基盤においてこの名称が見られるようになる。2008 年に米国人文科学基金（NEH）において DH への助成を専門的に扱う部署として Office of DH が設置されたことは象徴的な出来事であった。このように、DH という言葉は、学術政策に対する研究者コミュニティからの旗印という側面を有していた。一方で、この時期は、Web のインフラ化が進むとともに Web2.0 が提唱され始めるなど、一般の人々の情報共有環境までも激変させるような事態が生じていた。すなわち、この言葉が登場した時期と、この言葉が指すところの、人文学においてデジタル技術を応用して研究を発展させるという営み自体が大きく変化する時期とが重なっており、結果として、DH という言葉が登場した時期を以てこの種の研究の変化を語ることが世界各地で行われるようになった。

本稿が目指すのは、東アジアの人文学・DH におけるデジタルテキスト構造化の動向を中心とした状況を手がかりとしつつ、東アジアにおける DH の可能性を検討することである。したがって、上記のような状況を背景とした上で、デジタルテキストの構造化に関する動向を中心としつつその周辺状況をみていくことになる。デジタルテキストの構造化とは、デジタル化されたテキストデータの構造を人やコンピュータが分析し、研究に利用しやすいように何らかの手法に沿ってそれを記述することを指す。現在の DH は、単なる人文学のみにとどまらず、人文学から総合知や共創知へのルートの一つとして、そこからさらにどのような広がりを実現し得るかということについても期待が集まりつつあり、そこにおいて人文学の既存の成果が適切に活用されるためには適切な構造化がより重要となる。その基盤となる部分は、言葉としての DH が登場するよりも前から着々と取り組まれてきている。そこで、ここでは主に日本の人文学におけるデジタルテキストの構造化に関する動向を中心としつつ東アジアのこれまでの関連する状況についても概観し、その上で、そこから期待される可能性について検討したい。

1. 大まかな時期の区分

人文学においてデジタル技術を応用する場合、どちらかと言えば、最先端の機器やシステムがすぐに使われるというよりは、むしろ、いわゆるエンドユーザー・コンピューティングと言われるような、末端で低コストにその時点の情報インフラを利用するという状況を踏まえることが望ましく、新規の技術や規格が登場した時点というよりは、むしろ、ある程度普及した時期をおおまかにみていくことで状況を把握しやすくなるかもしれない。それゆえ、厳密な時期の区分は別の議論に譲るとして、ここでは、以下のように、Web2.0 以前をパーソナルコンピュータ（パソコン）が登場し普及する時期（パソコン時代）、Web が普及する時期（Web 時代）、Web2.0 や Unicode が普及する時期（Web2.0 時代）という風に分けてみたい。

1980 年代：パソコンの登場と普及
1990 年代：パソコンの普及と Web の登場
2000 年代：Web の普及
2010 年代：Web2.0 と Unicode の普及

パソコン時代よりも前にもデジタル技術は人文学において用いられており、ロベルト・ブサ神父によるトマス・アクィナスの電子索引が 1940 年代に着想されたことは、DH の嚆矢としてしばしば言及される。日本においても、1950 年代に設立された計量国語学会の論文誌を見ると、50 年代の終わりにはコンピュータを用いた研究が掲載されている。また、著作権の切れた文学作品等のデジタルテキストを無償で共有する Project Gutenberg が米国イリノイ大学で始まったのは 1971 年のことであった。しかしながら、この時期にはそれほどの広がりはみせない。人文学者の間で利用が広がるのは IBM PC が発売され世界にパソコンが普及する契機がもたらされた 1980 年代のことである。以下、筆者が 2021 年に執筆した論考[1] に拠りつつ、各年代の状況を大まかに振り返ってみたい。

2. 1980 年代の状況

1980 年代、個人が自分のパソコンにデータを入力できるという環境が登場し、各地で電子データの作成が行われた。とはいえ、当時のパソコンは、現在のものに比べると圧倒的に容量が小さく処理能力は遅く、人文学の場合には、数値データを扱う分野とテキスト資料を扱う分野において主にデータ作成が行われるようになった。管見の限りでは、数値計算、テキスト検索、目録や索引の作成等に用いられたようである。日本の仏教学においては、「個人用コンピューターによる索引」という発表が 1982 年に行われ、その後、1983 年には「インド仏教学とコンピューター」「コンピュータによる Uttarajjhāyā の韻律解析 II」「貝葉のコンピュータ処理」といった研究発表が日本の仏教学関連では最大の学会である日本印度学仏教学会において行われていた。1987 年にはこの学会が「イント学仏教学におけるコンピュータ利用」と題するシンポジウムを開催し、さらに 1989 年からは文部省科学研究費補助金研究成果公開促進費（データベース）を取得して印度学仏教学分野の論文書誌データベースの構築を開始した。

この時期の日本国内の状況については、2021 年に筆者が執筆した論考を一部改編しつつ述べてみたい。各地の研究機関による目録・書誌データベースが徐々に構築されるようになり、一方で、テキストデータベースの構築も徐々に進められていた。組織としての取り組みだけでなく個人の研究活動がコミュニティを形成し始める状況もみられており、とりわけ注目してお

1）永崎研宣.「哲学・思想研究における人文情報学の可能性」. 哲学・思想論叢, no. 39, 筑波大学哲学・思想学会, 2021 年 1 月, p. 95(44)-78(61).

きたいのは、1985 年 9 月 3 日、東京大学大型計算機センター集会室において第一回の研究会が開催された「哲学者のためのテキスト・データベース研究会」[2] である。この時には、デカルトの『省察』等、哲学研究のためのテキスト・データベースを構築するための基礎設計や実際についての発表が 5 件行われた。第 2 回にはプラトンやトーマス・マンも登場し、さらに第 3 回目からは「哲学者のための」という名称が外れ、広く人文・社会科学研究者のための「テキスト・データベース研究会」となっていった。英語名は「JACH（Japanese Association for Computer and Humanities）」とした旨、『JACH ニューズレター』第一号（1986 年 9 月 18 日刊）にて報告されている。おそらくは、米国 DH の基幹学会である Association for Computer and Humanities（ACH）を受けた名称のように見受けられ、開始段階では何らかの関係があった可能性も想定される。この頃の発表テーマとしては、テキスト・データベースの各種機能や OCR、人工知能があげられる一方、その楽しさや課題についての発表も行われていた。その後、研究会で扱われる話題は徐々に幅を広げていき、キーワードを拾い上げてみるなら、プラトン索引、科学論研究、解釈学、漢字字体、中世英語英文学、ウィトゲンシュタイン、国語調査、漢文テキスト、和漢文 OCR、LaTeX、美術史、デーヴァナーガリー文字、スタインベック、アラビア語テキスト、G.H. ミード、英日自動翻訳、国文学研究、源氏物語、タイ語テキスト、中世朝鮮語、チベット語、太宰治、西ドイツ中世史、といったものがみられた。ただし、1980 年代の段階では、デジタルテキストの構造的な記述についての議論はほとんどみられない。

1980 年代のテキストデータベースに関する動向としては、JACH に加えて 1987 年頃より運営されていた情報処理語学文学研究会にも注目しておきたい。こちらの英語名は JALLC（Japanese Association for Literary and Linguistic Computing）となっており、欧州の DH 学会の前身である ALLC（Association for Literary and Linguistic Computing）を受けたものと考えられる。JACH が哲学研究を中心に始まったのに対して、こちらは文学・言語学を中心に始まったもののようである。こちらでも、1980 年代には構造化に関する議論はほとんどみられず、各自が独自のフォーマットでデジタルテキストを作成していた可能性が考えられる。

研究者コミュニティとしては、1980 年代の末には、情報処理学会に人文科学とコンピュータ研究会が設立され、また、情報を組織化・知識化する研究のコミュニティとして情報知識学会が設立されて活動を開始した。情報知識学会ニューズレターの設立記念号では JACH の活動紹介が掲載されており、[3] 当時のこの種の活動の活発さがうかがえる。学会研究会等の研究者コミュニティが設立されるということは、すでに一定数の研究者がそれに取り組んでいるために可能なことであり、この種の取り組みが日本においてもすでに研究としてそれなりの広がりを見せていたことがうかがえる。

2）この研究会及び後継の JACH については、学位評価・授与機構の土屋俊先生に当時の資料をいただくことができたため、それに基づいていた記述となっている。感謝とともに記しておきたい。

3）坂井昭宏.「コンピュータのなかの古典―テキスト・データベース研究会（JACH）の研究活動―」. 情報知識学会ニューズレター, no. 1, 1988 年 9 月, pp. 11–13.

デジタルテキストの構造化に関する国際的な動向としては、1987 年にニューヨーク州のポキプシーにて初回の会合が開催された TEI（Text Encoding Initiative）ガイドラインがこの時期の重要な取り組みとして注目される。電子化テキストをよりよく作成し共有するためにポキプシーに集まった人文学者と情報学者達には、米国のみならず欧州や日本からの参加者も含まれていた。長い議論の末に、その成果はポキプシー原則としてまとめられ、これに沿った人文学電子テキスト資料のための構造化ガイドライン、すなわち、TEI ガイドラインが作成されることとなったのである。

人文学資料のためにテキスト資料を電子化し、それを何らかの有用な構造として作成しようとするということは、人文学がその活動において資料中に見出す内的な構造について十分に理解している必要がある。そして、人文学と一口に言っても様々な分野があり、テキスト資料を扱うものに限ったとしても、哲学、歴史学、文学をはじめ、それぞれが資料に対する多様なアプローチを含んでいる。それゆえに、電子化・構造化ガイドラインは、人文学における個々のアプローチを熟知する当事者たる人文学者達の手によって策定が進められる必要があった。このことは、人文学における電子化の方針を策定することが、特定ドメインを対象とした情報構築についての研究という、一つの研究領域を形成していくことをも意味した。それまでは徹底して紙媒体の制約を活かすことで展開されてきた研究資料が、その制約を離れた時にどのような形であり得るのか、そしてそれによって研究がどのように変化していくのか、ということは、思弁的な検討課題であってきたが、この TEI という営みを通じてそれを実践的に検証することが可能となった。

しかしながら、TEI の登場した時代は、未だコンピュータの利用が広く普及していたわけではなくコンピュータ利用自体の技術的制約も強く、これがその意義を踏まえて大きく花開くには、今少しの時間を必要とした。

以上、1980 年代の状況を大まかに見てきた。この流れは、ほぼそのまま 1990 年代に引き継がれていく。

3. 1990 年代の状況

1990 年代についても、永崎 2021 に依拠しつつみてみよう。国内状況としては、この時期、テキストデータベースや電子テキストを主に扱ってきた JACH・JALLC は活動を発展させていく。

JACH の 1991 年 6 月 28 日の第 13 回研究会では、当時 TEI を推進していた研究者の一人であるオックスフォード大学の Susan Hocky 氏による講演ともに TEI が主に取り上げられサンスクリット・文学研究・日本文学研究・マルチメディアへの適用可能性が議論されたようである。筆者が入手した資料ではこの次の 1991 年 12 月 13 日の仙台国際センターでの第 14 回研究会の開催を最後に記録が途絶えている。この会では、インド学仏教学からコンピュータに取り

組む 6 件の発表と TEI に関する発表が 1 件という構成であり、その様子が情報知識学会ニューズレターで伝えられている。[4)]

JALLC においても、国際的な動向との連携についての議論が行われたようであり、TEI に関する発表が散見され、検討が行われたことは見てとれるが、一方で、文字コードの問題や SGML 時代の TEI の課題なども指摘されており、[5), 6)] そもそも OS が多言語対応できず国際的な成果の共有が困難であったことなどから、当時の日本で導入することが容易ではなかったことが想定される。また、1993 年のシンポジウム報告[7)] では、JALLC の主な活動が文学作品等の電子テキストの会員同士での流通という互助的なものであることを示す一方で、一般公開の条件については 1989 年に米国で始まった GNU GPL（General Public License）によるコピーレフト運動を検討の俎上に載せており、現在のオープンサイエンス・オープンデータ推進の先駆的な活動という側面もみてとれる。なお、JALLC の活動は、筆者が確認できた限りでは 1998 年 7 月 18 日を最後としている。このような人文学発の研究者コミュニティによる自発的な取り組みは、ここで一度途絶えてしまう。

国際的な動向との連携としては、上述の JACH や JALLC の活動にみられるように、できたばかりの TEI ガイドラインを日本にも導入しようとする動きがあったようである。このことは、上述の情報知識学会のニューズレターにおいてもその片鱗を垣間見ることができる。その後、哲学研究者としてテキスト・データベースに熱心に取り組んでいた長瀬真理氏により TEI に関する報告も行われていた。[8)]

また、研究集会としては、京都大学大型計算機センターの研究セミナーとして 1990 年 3 月に初回が開催された「東洋学へのコンピュータ利用」研究セミナーも挙げておきたい。これはテキストデータベースというよりはデータベース全般ということで東洋学におけるコンピュータ利用に関する様々な取り組みが毎年報告されてきており、これは現在でも毎年一度、活発な研究会が開催されている。

この頃は、テキストデータベース構築のプロジェクトも各地で推進されるようになっていた。国文学研究資料館では、日本古典作品の校訂本文データベースの構築のみならず、その構造化を行うためのルールの策定も行われた。[9)] その後このルールは KOKIN ルールと名付けられ、汎

4）三木邦広.「JACH 第 14 回研究会に参加して」. 情報知識学会ニューズレター, no. 13, 1992 年 4 月 1 日, pp. 2–4. https://www.jsik.jp/archive/news/N13.pdf#page=2

5）豊島正之.「TEI からみた SGML のはなし」. 情報処理語学文学研究会報, no. 12, 1992 年 12 月, https://www.joao-roiz.jp/mtoyo/TEI/JALLC-12-TEI.pdf.

6）「TEI-P3 について」. 情報処理語学文学研究会報, no. 15, 1994 年 7 月, https://www.joao-roiz.jp/mtoyo/TEI/JALLC-TEIP3.pdf.

7）内田保廣.「『情報処理語学文学研究会』のテキスト・アーカイブス」. 情報知識学会誌, vol. 3, no. 1, 1993 年 12 月, pp. 45–51.

8）長瀬真理.「TEI の活動と今後の展望」. 情報知識学会ニューズレター, no. 10, 1991 年 10 月, pp. 9–10.

9）安永尚志.「日本古典文学本文データベース形成とデータ記述文法」. 情報処理学会「人文科学とコンピュータ」研究報告 1991, no. 20（1990–CH–008）(1991 年)：1–8.

用性を高めるべく、SGML 化が行われ、[10] TEI との互換性も検討課題に挙げられていた。[11] また、フッサールやヘーゲル[12] のテキスト・データベースが日本で作成され、関係者の間で活用されたようだが、版権問題で公開が十分にできないこと[13] もあったようである。宗教関連文献でも枚挙に暇がないが、国内の例を挙げてみるなら、たとえばサンスクリット仏典の全文テキストデータベースに基づいて作成された「ダルマキールティ梵文テクスト KWIC 索引」[14] では刊行後に元になったテキストデータベースも Web 公開された。あるいはまた、パーリ語聖典のデータベース化への取り組み、[15] 1994 年の SAT 大蔵経テキストデータベース研究会（代表：下田正弘・東京大学大学院人文社会系研究科教授）による大正新脩大藏經への着手など、[16] この頃には大規模なものへの組織的な取り組みも行われるようになっていた。さらに、青空文庫が 1997 年に開始され、著作権切れのテキストデータが自由に利用可能になったことも、この時期の重要な取り組みとして位置づけられるだろう。

このような流れの終わりの時期にあたる 1998 年には漢字文献情報処理研究会が設立された。同研究会はこの年に『電脳中国学』[17] を刊行し、中国学における当時のデジタル技術の状況を伝えてくれている。ここでは台湾中央研究院の漢籍全文資料庫をはじめとして中国語学のためのデジタルリソースやその活用の仕方等が紹介されており、様々な組織や個人が研究リソースの提供を開始していることがうかがえる。また、当時は Web だけでなく CD-ROM によるデータベース販売も広く行われていたようである。

その一方で、人文学における研究基盤の構築のための要として特に欧米ではその後着々と普及していった TEI については、日本では導入に向けた動きの後、しばらくの間、おそらくは上述の理由により、徐々に姿が見えなくなってしまう。

TEI の国際的な動向としては、1990 年には TEI ガイドラインの第一版である TEI P1、1994 年には TEI P3 が発行され比較的安定的なものとなる。これらは、SGML というマークアップ言語を用いて人文学資料における構造を記述するというものであった。さらに、1998 年には、ここで得たマークアップ言語の知見が部分的に反映される形で、XML（Extensible Markup

10）安永尚志.「国文学作品のテキストデータ記述ルールについて」. 自然言語処理 3, no. 4（1996 年）: 3–29. https://doi.org/10.5715/jnlp. 3.4_3.
11）原正一郎, 安永尚志.「国文学研究支援のための SGML/XMS データシステム—国文学データ共有のための標準化—」. 情報知識学会誌 11, no. 4（2002 年）: 17–35, 46. https://doi.org/10.2964/jsik_KJ00001039453.
12）ヘーゲル・テキストデータベース　https://www.unii.ac.jp/iori/jheg1.html（2022–01–10 確認）
13）浜渦辰二.「峠を越えたフッサール・データベース：インターネット時代のマルチリンガル・テキストのために」. 人文論集, vol. 48, no. 1, 1997 年 7 月, pp. A1–29.
14）小野基・小田淳一・高島淳「ダルマキールティ梵文テクスト KWIC 索引」, 東京外国語大学アジア・アフリカ言語文化研究所 1996.3.15. ISBN 4–87297–444–1
15）中谷英明, 江島恵教.「パーリ三蔵データベースの構築と仏典研究」. パーリ学仏教文化学, vol. 8, 1995 年, pp. 123–47, doi:10.20769/jpbs. 8.0_123.
16）下田正弘, 永崎研宣. デジタル学術空間の作り方：仏教学から提起する次世代人文学のモデル. 文学通信, 2019 年.
17）漢字文献情報処理研究会編『電脳中国学』好文出版, 1998 年.

Language, 拡張可能マークアップ言語）が World Wide Web Consortium（W3C）で制定された。XML は人文学のみならず、情報技術の世界全般に大いに利用されることになり、現在では Microsoft の Word や Excel、PowerPoint もこの XML を採用している。

また、個別分野の動向として一例を挙げておくと、1992 年、仏典テキストの電子化と利用についての連携の場として EBTI（Electronic Buddhist Text Initiative）が活動を開始し、世界中の仏教研究のデジタル化に関するプロジェクトの交流の場として活動していた。

1990 年代の後半には、文字コードの問題を解決するソリューションが登場してくる。京都大学人文科学研究所における e 漢字、今昔文字鏡研究会における今昔文字鏡フォント、東京大学と日本学術振興会による GT フォント、等がそれにあたる。いずれも、既存の文字コードでは表現しきれない文字や字形を表示させるべく、数万字の文字を表示するための手立てであり、そのような細やかなニーズを持つ人文学研究者がいよいよテキストデータの構築に踏み込んできたことをうかがわせる事態であった。ただ、いずれの事業も、2021 年現在は活動が見えない状況となっており、その一方で、Unicode でカバーできる文字の範囲が圧倒的に広がっている。当時の不足を補うアドホックな取り組みとして理解しておくのがよいだろう。

4. 2000 年代の状況

2000 年代に入ると、Windows2000 やその後に大ヒットする WindowsXP において Unicode が全面的に採用され、Unicode が多言語利用のための現実的な選択肢となっていく。Web はインフラストラクチャーと言えるような普及の段階になり、さらに 2005 年には Google Map が登場したことで、インターネットに接続されたパソコンがあれば Web で地図上に共同でデータ構築を行うといったやや複雑なことでさえも誰でも取り組めるようになった。すなわち、国境や文字種の壁を越えたコンテンツの即時的な相互運用が容易に手の届くものとなったのである。このような流れのなかで、上述のように、DH が登場することになる。この時期のデジタルテキストをめぐる状況について少しみてみよう。

2001 年に刊行された漢字文献情報処理研究会による『電脳中国学 II』では、Unicode を採用した Windows2000 により、中国語と日本語のコンテンツを横断的に利用可能な多漢字環境が提供されるようになったことを伝えている。同時に、今昔文字鏡が 9 万字を提供しているという記事も掲載されるなど、未だ十分な多漢字環境は提供されていないにせよ、Web によって形成されたグローバルな規模のハイパーテキストの網目にとって、単一の文字コードですべてのテキストを処理できる仕組みは大きなインパクトをもたらしたようであることがうかがえる。一方で、本書では、テキスト構造化に関して日本が諸外国に後れをとっていることについての危機感も述べられている。

文字の扱いに関しては、上記のような動向に対して文字コードを利用すること自体の本質的な問題も提起され、文字素性に基づく文字処理を実現する CHISE（Character Information

Service Environment）プロジェクトが開始されている。[18)]

また、韓国において高麗大蔵経研究所、台湾においては中華仏典協会（CBETA）、日本でもSAT大蔵経テキストデータベース研究会が、相次いで大蔵経のデータベースを完成・公開し、仏典の大規模な横断テキスト検索等が可能となり、テキストデータベースはより身近なものになっていった。

国際的な状況としては、テキストデータベースのみならずDH全体の理念的背景として広く共有される概念であるMethodological Commons（方法論の共有地）がキングスカレッジ・ロンドンの研究者らによって最初に発表されたのは2002年のことである。人文学における多様な方法論がデジタル技術を通じて相互に連携し、よりよいデジタル技術の活用手法を追求するとともに自らの方法論を反省する機会とし、それを通じて新たな研究成果を産み出していく場を形成することがデジタル時代の人文学においては重要であり、それが数年後に言葉として登場するDHにおいても理念的背景となった。これは、人文学全般にわたる国際的なデジタルテキストの構造化を目指すTEIガイドラインと親和性の高いものであった。

DH専門の助成金が本格的に設定されるようになり、欧州でもDH向けの研究助成金が広く配分され、二国間[19)]や複数国間[20)]での共同助成金も開始されるなど、研究助成の側からもDHが強く支持されるようになり、これを踏まえた人文学研究基盤の整備も着実に進められることとなる。ここでも上述のTEIをはじめとして、それまでに策定されてきた様々な文化資料向けの規格が、そうした研究助成の期待に対応できる堅実な受け皿としての力を発揮することになる。

一方、このような流れの中で、デジタルテキストを用いる研究では2つの方向性が明確になっていく。一つは、これまでは紙のテキストに暗黙的に込められてきた、人による理解や解釈の結果をより深く研究対象として共有・活用するという方向性と、増大し続ける分量と言語的多様性のために精読することが難しいテキスト群をコンピュータで処理・分析するという方向性である。前者については、TEIガイドラインのP5が2007年にリリースされた。[21)] 後者のための方法論としては、遠読（Distant reading）[22)]が提唱されるようになった。以下、この2つについてそれぞれ簡単にみてみよう。

TEIガイドラインはP4からXMLには対応していたものの、P4はSGML向けに作成されたP3を継承しつつXMLに対応させたものであり、XMLの特性を活かしたものとはなって

18）守岡知彦，師茂樹．「文字素性に基づく文字処理」．情報処理学会研究報告．CH,［人文科学とコンピュータ］62（2004年5月）：53-60.

19）NEH/DFG Bilateral Digital Humanities Program https://www.federalgrants.com/NEH-DFG-Bilateral-Digital-Humanities-Program-36894.html（2022-01-10確認）

20）Digging into Data Challenge https://diggingintodata.org/（2022-01-10確認）

21）Burnard, Lou, Syd Bauman. The TEI Guidelines. TEI P5 Guidelines, 2007年．https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html.（2022-01-10確認）

22）Moretti Franco. 秋草俊一郎他共訳．遠読：「世界文学システム」への挑戦．みすず書房，2016年.

いなかった。それを解決した新しいバージョンが P5 であった。これ以降、TEI ガイドラインは P5 をさらに改訂するという形で発展することになる。

遠読は、世界文学研究の現場から生じてきたものであり、[23] 様々なコンピュータのツールを用いた統計分析を通じてテキストの特徴を把握することによってテキストに内在するものを理解しようとする研究動向である。これは、Google Books 及びそのデータを大学図書館が主体となって扱うべく設立された大学図書館連合による電子図書館 HathiTrust Digital Library において、蓄積された膨大なテキストデータを分析するための考え方として用いられる[24] など、ビッグデータ時代における DH の可能性を示すものとして受容されるようになっていった。

5. 2010 年代の状況

2010 年代に入ると、DH は国際的にもさらなる広がりをみせ、2012 年の日本 DH 学会（JADH）の設立と ADHO への加盟をはじめとして、世界各地で DH 学会が設立され、ADHO へと加盟していくことになる。DH が志向するグローバル性が、中核としてのグローバルな組織の主導よりもむしろローカル同士の丁寧な連帯によって成立するという状況が徐々に共有されていく時期であったと言えるだろう。この時期は、日本以外にも、オーストラリア圏、フランス語圏、メキシコ、南アフリカ共和国、台湾が加盟[25] し、さらに、欧州 DH 学会[26] 内でドイツ語圏やイタリア、チェコ、北欧、ロシアの地域学会が設立されそれぞれに学術大会を開催するなどの活動を開始している。

そのような流れの一方で、DH においてはクラウドソーシングという形での市民との連携に期待が集まるようになる。一例として、Transcribe Bentham プロジェクトをみてみよう。[27] このプロジェクトは、元々、著名な思想家である Jeremy Bentham の草稿を翻刻（文字起し）して出版するプロジェクトであったが、これが DH の力を借りてクラウドソーシングによる翻刻を実施したことにより、国際的にも多くの参加者を集め、それまでとは比較にならない大量のテキストデータを短期間で作成でき、注目を集めることになった。DH におけるこの種のプ

23）なお、遠読と呼び得る研究手法はコンピュータを利用せずとも行われていたという見解もある。Cf. Underwood, Ted.「A Genealogy of Distant Reading」. Digital Humanities Quarterly, vol. 011, no. 2, 2017 年 6 月.

24）こちらのリストを参照されたい。https://www.hathitrust.org/usage-examples-of-hathitrust-datasets（2022-01-10 確認）

25）https://adho.org/（2022-01-10 確認）

26）https://eadh.org/（2022-01-10 確認）

27）Terras, Melissa. Present, not voting: Digital Humanities in the Panopticon: closing plenary speech, Digital Humanities 2010. Literary and Linguistic Computing, vol. 26, no. 3, 2011年, pp. 257-69, doi:10.1093/llc/fqr016. この論文の元になった講演録の児玉聡氏による日本語訳は以下の URL で参照できる。https://www.dhii.jp/dh/dh2010/DH2010_Plenary_trans_by_kodama.html（2022-01-10 確認）

ロジェクトとしては初期の大きな成功であり、その結果、様々な研究助成金を得て[28] さらに発展することとなった。世界中で多くの研究者が専門的に取り組んでいる Bentham ならではのことではあり、同様のことがすぐに他にも適用できるわけではないが、事例としては大いに参考になる。日本でも 2017 年には京都大学古地震研究会により「みんなで翻刻」が公開され、クラウドソーシング翻刻が本格的に行われるようになった。

同様の状況は TEI のコミュニティにおいても見られるようになる。そもそも上述のように、TEI ガイドラインは、コミュニティに集った研究者達のニーズにあわせて策定されてきたものであり、欧米の研究者しかそこにいなければ欧米の資料が対象の中心にならざるを得ない。TEI ガイドラインには外字に関する項目もあり、そこではすでに東アジア研究者による貢献もあった。ただし、外字は、大量の漢字を扱わねばならず、Unicode に登録されていない漢字としての外字も様々に扱わねばならない漢字文化圏の宿痾であり、同時にそれが一つの特徴であるとも考えられがちだが、欧州においても中世写本に記載された多くの特殊な文字の表現に苦労してきた一面があり、[29] 外字はむしろグローバルな課題の一つとして捉えられてきたとみることができる。

グローバルとローカルの関係において注目されるのが、TEI 協会における 2016 年の東アジア／日本語分科会（SIG East Asian/Japanese）の設立である。TEI ガイドラインは当初より、一般的なテキスト性を踏まえた資料の構造化を目指してきていたと見ることができるが、この分科会の設立は、そのような方向性に対して、むしろ個々のローカルな言語文化圏に蓄積されてきた慣習、いわばテキスト伝統に対応するという姿勢を示すものだったということができる。東アジア／日本語分科会は、かつて行われた TEI ガイドラインの用語説明の日本語訳[30] を踏まえた TEI ガイドラインの翻訳や日本語資料向けガイドライン[31] の作成、そしてそれらを踏まえた TEI ガイドライン自体の再検討などの取り組みを開始した。これに続いて、漢字文化圏とは異なるテキスト伝統を有するインドテキストの分科会も設立されることとなった。さらに、2019 年には国際化ワーキンググループも設置され、多言語グロッサリーの開発やガイドライン翻訳の簡易な手法の確立と実践を目指して活動を行っている。このこともまさに、ローカル同士の丁寧な連帯によってグローバル性の実現を目指す典型的な例であると言えるだろう。

現在、デジタルテキストの扱いについては、作成・構造化・処理等のいずれの過程においても Unicode で文字が符号化されていなければかなり非効率的になってしまう。この時期には、Unicode で使用可能な文字が飛躍的に増加しただけでなく、Unicode では同定されてしまう微細な漢字の字形差についてもテキストデータのレベルで記述できるようにする IVS（Ideographic Variation Sequence）が普及し、10 万種以上の漢字字形が Unicode で利用可能となった。そ

28）Funding https://blogs.ucl.ac.uk/transcribe-bentham/funding/（2022-01-10 確認）
29）https://mufi.info/（2022-01-10 確認）
30）これは、TEI P5 の初期バージョンについて、鶴見大学の大屋一志先生によって行われた翻訳である。
31）https://github.com/TEI-EAJ/jp_guidelines/wiki（2022-01-10 確認）

れだけでなく、変体仮名や悉曇も異体字まで利用可能となり、Unicode を基盤とする学術研究環境はいよいよ充実することとなった。[32)]

この時期には、人文学における画像の扱いについても画期が訪れる。2011 年頃より世界の有力な研究図書館が連携して IIIF（International Image Interoperability Framework）という画像共有方式を開発し、これが世界中の研究図書館のデジタル画像公開において用いられるようになった。2010 年代には国内でも国立国会図書館や国文学研究資料館をはじめ様々な資料所蔵機関から IIIF に準拠した画像が公開され、その相互運用性の高さにより多様な活用が行われるようになっている。[33)] テキストデータとの連係も行われたが、構造化デジタルテキストとの相性がよく、TEI ガイドラインに準拠したテキストと IIIF 準拠画像をリンクさせて利便性を高める取り組みが世界各地で進められた。

この時期の中国学に関して注目すべきイベントとして、2018 年に上海で開催され、世界中の中国史に関するデジタル研究プロジェクトが一堂に会した International Conference on Cyberinfrastructure for Historical China Studies[34)] がある。とりわけ活発のように思われたのは、ハーバード大学が取り組む人物データベース「CBDB（China Biographical Database Project）」、[35)] ライデン大学で推進される固有名詞自動タグ付けシステム「MARKUS」、中国語テキストクラウドソーシング共有システム「中國哲學書電子化計劃（Chinese Text Project）」[36)] であり、MARKUS はテキスト中の固有名詞をタグ付けした情報の共有に際して TEI に準拠した形式を採用していた。

6. 2020 年代から今後の可能性へ

ここまでの状況及び 2020 年代に入ってからのいくつかの事柄を踏まえ、今後の可能性について検討してみたい。2020 年代は、その始まりとともに COVID-19 に翻弄され、本稿執筆時点では未だに収束の見通しはない。DH もまた、対面での研究集会ができず交流が弱まってしまう面はあったものの、逆に国際的な集会をオンラインで開催することについてのコンセンサスを得やすくなり、これを踏まえた活動が着実に進められてきた。2020 年 3 月にリリースされた Unicode 13.0 では、CJK 統合漢字拡張 G が含まれることとなり、9 万字を超える漢字が Unicode で利用できることとなった。また、2021 年 2 月には、TEI ガイドライン P5 4.2.0 がリリースされた際に日本語で主に用いられるルビのルールが組み込まれた。これは東アジア／

32）なお、この一連のプロセスにおいては、TEI 協会東アジア／日本語分科会の設立や、学術団体が主体となって漢字を Unicode で符号化すること、これに加えて悉曇の異体字導入については、大正新脩大蔵経デジタル版の高度化に取り組む SAT 大蔵経データベース研究会により実現されたことは特筆しておきたい。

33）国内の活用事例については https://digitalnagasaki.hatenablog.com/iiif を参照されたい。（2022-01-10 確認）

34）https://ctext.org/digital-humanities/shanghai2018（2022-01-10 確認）

35）https://projects.iq.harvard.edu/cbdb/home（2022-01-10 確認）

36）https://ctext.org/zh（2022-01-10 確認）

日本語分科会の活動の重要な成果と言えるだろう。これにあたっては、世界各地の他のローカルなルールをこれまで導入してこなかったこととどのように整合させるのかという問題提起もなされたが、どれくらい一般化可能かという議論を個々の事例に応じて積み重ねていくことになるようだ。このようにして、欧米のテキストには用いられないルールが TEI ガイドラインに導入されたことは、今後、このような流れが東アジアのみならず、世界に開かれていき、グローバルな視点からより適切に通用するものとなっていくことをも意味しており、東アジア／日本語分科会がその先鞭をつけたと言うこともできるだろう。

また、ここに至り、日本や東アジアのデジタルテキストの構造化において TEI ガイドラインを採用したデジタル化プロジェクトが日本でも公表されるようになってきた。国文学研究資料館における日本古典籍の試行的 TEI マークアップ、[37] 国立歴史民俗博物館における延喜式への歴史史料としてのマークアップ、[38] SAT 大蔵経データベース研究会の SAT2018 における現代日本語訳仏典や漢文仏典、[39), 40), 41)] 渋沢栄一記念財団による『渋沢ダイアリー』における日記へのマークアップ、[42] 関西大学アジア・オープン・リサーチセンターによる『廣瀬本万葉集』への左右の訓等にも対応した詳細な写本情報のマークアップ、[43] 有志団体による『デジタル源氏物語』[44] における現代語訳と古典の対訳マークアップとのリンクなどがあり、これらの多くは IIIF 対応画像とリンクされており、Web インターフェイス上ではテキストを読みながら対応する画像を閲覧できるようになっている。また、国際的な協働プロジェクトとして、2021 年に公開された、フランス学士院碑文・文芸アカデミー（Académie des inscriptions et belles-lettres）と SAT 大蔵経データベース研究会の協働の成果である Digital 法寶義林[45] では、『法寶義林』の人名索引を既存のリソースや関連する地理座標等とリンクする形で TEI マーク

37）この成果は https://github.com/TEI-EAJ/jpn_classical（2022-01-10 確認）において事例として掲載されている。

38）小風尚樹，中村覚，永崎研宣，渡辺美紗子，戸村美月，小風綾乃，清武雄二，後藤真，小倉慈司．相互運用性を高めた日本歴史資料データ実装：『延喜式』TEI と IIIF を事例として．じんもんこん 2021 論文集，no. 2021（2021 年 12 月）：294-300.

39）渡邉要一郎，永崎研宣，朴賢珍，王一凡，村瀬友洋，渡邉眞儀，大向一輝と下田正弘．「大正新脩大蔵経の構造的記述に向けて」．じんもんこん 2020 論文集，no. 2020（2020 年 12 月）：61-66.

40）王一凡，渡邉要一郎，永崎研宣，下田正弘．『續一切經音義』からみる漢文文献の TEI マークアップの課題．じんもんこん 2021 論文集，no. 2021（2021 年 12 月）：234-239.

41）左藤仁宏，渡邉要一郎，永崎研宣，下田正弘．仏教思想の概念体系の記述手法としての TEI マークアップの現状と課題．じんもんこん 2021 論文集，no. 2021（2021 年 12 月）：288-293.

42）金甫榮，中村覚，小風尚樹，橋本雄太，井上さやか，茂原暢 , 永崎研宣．「TEI を用いた『渋沢栄一伝記資料』テキストデータの再構築」．じんもんこん 2020 論文集，no. 2020（2020 年 12 月）：47-52.

43）永崎研宣，ほか．万葉集伝本研究のためのデジタル基盤構築 廣瀬本『万葉集』の構造化とビューワの開発．2，一般財団法人人文情報学研究所，関西大学，関西大学，関西大学，合同会社 AMANE，合同会社 AMANE，佐賀大学，2021 年 2 月．

44）サイトについては、https://genji.dl.itc.u-tokyo.ac.jp/（2022-01-10 確認）．中村覚 , 田村隆，永崎研宣．「源氏物語本文研究支援システム『デジタル源氏物語』の開発における IIIF・TEI の活用」．研究報告人文科学とコンピュータ（CH）2020-CH-124，no. 2（2020 年 8 月 29 日）：1-7.

45）Digital 法寶義林 https://tripitaka.l.u-tokyo.ac.jp/hbgrn/（2022-01-10 確認）

アップを行ってオープンライセンスで公開し、さらにそのデータを用いた検索やそれを地図・年表上に表示するといったサービスを提供している。このように、日本における TEI に準拠した構造化にも様々なアプローチが登場してきている。そして、国文学研究資料館等の日本語古典籍のマークアップの成果は、日本語古典籍 TEI 本文データ作成要領[46] としてまとめられ、前出の日本語資料向け TEI ガイドラインを構成するものとして Web で公開されている。さらに、テキスト構造化を支援する環境として、関西大学アジア・オープン・リサーチセンターにより、東アジア DH ポータル[47] が公開され、主に欧米の DH において利用されてきたデジタルツールやメソッドが積極的に翻訳・紹介されており、なかには TEI ガイドラインの一部の翻訳も含まれている。[48]

このように、デジタルテキストの構造化については、日本でもようやく本格化しつつあり、まだ課題は山積しているものの、そういった課題を解決していく上で必要となる、東アジア・日本語資料のための TEI ガイドラインの改定や、漢字を含む学術目的の文字の Unicode における符号化のための体制が整ったと言える状況にある。すなわち、日本を含む東アジアにおけるテキスト資料に関して、どのような構造が共有されるべきなのか、それはどういう意味があるのか、といったことにはじまり、より抽象度の高い事項に至るまで、構造化に必要と考えられる要素を、国際的な標準に反映させることを一連のプロセスの中の選択肢の一つとしながらグローバルに共有・活用できる環境が整備されたのである。このプロセスに東アジアの人文学者がより参画しやすくなる手立てを今後用意する必要はあるものの、ここにきてようやく、欧米の DH とほぼ同等の環境にたどり着いたということもできるだろう。

一方で、IIIF の普及にもみられるように、デジタル画像をはじめとするマルチメディアデータに関しては言語文化による差異がテキストデータに比べると少なく、現時点では日本の取り組みも国際的には先進的な部類に入ると言ってもよい状況にある。これに構造化テキストをうまく組み合わせることができれば、東アジアの DH の特徴を活かした研究活動を推進できるだろう。デジタルテキストの東アジア的な構造化の取り組みに加えて、マルチメディアデータとの連携により東アジアの DH がもたらし得る多様性は、グローバルな DH をも裨益し、ひいては人文学そのものをよりよいものにしていくことに貢献することになるだろう。

46）日本語古典籍 TEI 本文データ作成要領　https://github.com/TEI-EAJ/jpn_classical/blob/master/jpn_classical_guideline.md（2022-01-10 確認）
47）東アジア DH ポータル　https://www.dh.ku-orcas.kansai-u.ac.jp/（2022-01-10 確認）
48）菊池信彦，川創，二ノ宮聡.「『東アジア DH ポータル』の構築と課題：デジタルヒューマニティーズの研究ノウハウのオープンな知識基盤を目指して」. じんもんこん 2020 論文集 2020（2020 年 12 月）：229-34.