人文情報学月報第 147 号【前編】
Digital Humanities Monthly No. 147-1
ISSN 2189-1621 / 2011 年 08 月 27 日創刊
2023 年 10 月 31 日発行 発行数 1092 部

《巻頭言》古辞書研究と DH にまつわる体験からみえたこと

（李 媛：京都大学人文科学研究所附属人文情報学創新センター助教）

筆者は 2023 年 10 月 14 日に羽田空港を後にし、アメリカのイェール大学で 2023 年 10 月 16 日から 20 日まで開催される IRG （Ideographic Research Group） #61[1]の会議に参加するために向かった。これは、Unicode の対応規格である ISO/IEC 10646 において漢字の符号化を検討する IRG の定例会議に、ISO/IEC JTS1/SC2 委員会のリエゾンメンバーとして同会議に正式に参加している唯一の学術団体である SAT 大蔵経データベース研究会のメンバーとして参加するとともに、符号化を提案された漢字のレビューの議論に参加するためであった。今回、筆者が IRG 会議に参加するのは 5 回目となった。コロナの影響で前の 4 回はオンラインでの参加だったが、今回は初めて対面での参加となった。本稿は、この会議の合間を縫って、これまでの筆者の古辞書研究と DH（デジタル・ヒューマニティーズ）の経験を振り返る機会として執筆するものである。

2011 年 4 月から北海道大学文学部研究生として、筆者は言語情報学講座の研究室に配属されることになった。それまでの人生では、学部時代はコンピュータ情報処理を専攻し、大学を卒業した後、しばらくは民間企業に勤務していた。一方で、人文学、特に漢字研究に魅かれ、京都での一年間の交換留学中に経験した日本語の漢字と中国の簡体字との間の違いに強く関心を持つようになった。そして、この体験によって「異郷有悟」という思いが心に深く残り、北海道大学の日本古辞書研究を専門する国語学者である池田証壽先生の研究室を志望して入学を決意した。

研究生として入学した際の初めの研究計画は簡体字をテーマとするものだった。これは、自分の中で長らく持っていた簡化字と繁体字とのギャップという素朴な問題意識から生まれたものだった。しかし、もし博士課程を目指すのであれば、簡体字の歴史は比較的浅いため、より深く漢字研究を進める方が良いと先生からの提案があり、日本に現存する最古の漢字字書高山寺本『篆隷万象名義』を研究対象とすることになった。

最初に、『篆隷万象名義』の影印テキスト、そして大漢和辞典の番号と漢字の対応するエクセルファイルのリストを渡されて、そこから『篆隷万象名義』の全文テキスト化作業を続けることを提案された。後に知ることとなったが、池田先生はすでに 1994 年頃から、『篆隷万象名義』のテキスト化の取り組みを始めていた[2]。当時、パソコンの漢字処理環境は初歩的な段階だった。特に古い文献の処理においては漢字不足の問題が際立っていた。しかし、池田先生は将来、漢字処理の課題が克服されるとの確信を持っており、国語学者として、

多くの難字を含む古辞書のデジタル化への挑戦を開始した。そして、日本の漢字コード化である JIS 漢字コードの取り組みにも寄与していた。実際に、日本の人文学における DH の研究の進展は、大規模漢字符号化集合 Unicode の整備に大きく影響を受けていると考えられる（表 1、[4]）。

表 1 Unicode and digital humanities in Japan

Year	Size of character set	Standard	CJK Unified Ideographs	Number of Chinese characters	PC operating system	Digital humanities in Japan
1963		ASCII		0		
1978	Around 6,000	JIS C 6226-1978		6,349		
1981					MS-DOS 1.0	
1983		JIS X 0208-1983		6,353		
1984					MS-DOS 3.0	
1988						YDIC[1]
1990		JIS X 0208-1990		6,355		IPSJ SIG CH[2]
		JIS X 0212-1990		5,801		
1991		Unicode ver. 1.0		0		
1992	Around 20,000	Unicode ver. 1.01	URO	20,902		JALLC[3]
1994						**KTB — Temporary version (entries)**
1995					Windows 95	
1996		Unicode ver. 2.0	URO	20,902		
1997		JIS X 0208:1997		6,355		
1998						JAET[4]
1999		Unicode ver. 3.0	URO, Extension A	27,484		JINMONKON[5]
2000		JIS X 0213:2000		3,685	Windows 2000	
2001	Around 70,000	Unicode ver. 3.1	URO, Extension A-B	70,195	Windows XP; MacOS X	
2003						Kanji DB[6]; **KTB — Mojikyo version (entries)**
2004						HNG[7]
2005		Unicode ver. 4.1	URO, Extension A-B	70,217		Takuhon-moji Database[8]; CHISE IDS[9]
2006						GlyphWiki[10]
2007						SAT[11]
2008		Unicode ver. 5.1	URO, Extension A-B	70,225		
2009		Unicode ver. 5.2	URO, Extension A-C	74,374	Windows 7	
2010		Unicode ver. 6.0	URO, Extension A-D	74,596		
2011						**KTB — UCS version (entries)**
2012		JIS X 0208:2012			Windows 8	CHJ[12]
2014						NIJL-NW[13]; **HDIC Established**
2015	Around 80,000	Unicode ver. 8.0	URO, Extension A-E	80,358		
2016		Unicode ver. 9.0		80,358		**KTB — UCS version (full-text)**
2017		Unicode ver. 10.0	URO, Extension A-F	87,861		

[1] Masayuki Toyoshima, Hokkaido University (present affiliation is Sophia University); [2] IPSJ SIG Computers and the Humanities; [3] Japanese Association for Literary and Linguistic Computing; [4] Japan Association for East Asian Text Processing; [5] Information Processing Society of Japan, Special Interest Groups, Computers and the Humanities; [6] Taichi Kawabata, NTT; [7] Hanzi Normative Glyphs, Harumichi Ishizuka, Hokkaido University; [8] Character Database of Digital Rubbings, Koichi Yasuoka, Kyoto University; [9] Character Information Service Environment, Tomohiko Morioka, Kyoto University; [10] Koichi Kamichi, Keio University (present affiliation is Daito Bunka University); [11] The SAT Daizōkyō Text Database Committee, University of Tokyo; [12] Corpus of Historical Japanese, National Institute for Japanese Language and Linguistics (NINJAL); [13] Project to Build an International Collaborative Research Network for Pre-modern Japanese Texts, National Institute of Japanese Literature (NIJL).

しばらくの間は、このような作業を続けた。はじめのうちは、古文献に触れる経験がない上に、難解な漢字をコンピュータ上に文字コードで表現することも未経験だったため、どのようにしてそれが研究の内容と関連しているのかを把握するのが難しく、挑戦的な日々だった。後になって理解できるようになったが、この感覚は初心者の段階でのものだった。古辞書に関する授業を受講する中で、関連する文献を参照しつつ、『篆隷万象名義』の本文解

読やテキスト化作業を少しずつ進めていった。

1994 年に、高山寺本篆隷万象名義データベース（Kosanjibon Tenrei Bansho Meigi database, 略称 KTB）の掲出字の電子化を始めた際、Unicode ver. 1.01 では 20,902 字の漢字が扱えた。その時点で、KTB の掲出字テキスト化のテンポラリバージョンが作成され、符号化できなかった部分は 69.1%（第一〜四帖、目録部分を含む）であった[2]。しかし、その後の CJK 統合漢字[3] Ext A-B の追加により、Unicode ver. 3.1 では扱える漢字の字数は 70,195 字に増加した。それにもかかわらず、この時点では Unicode がパソコンに実装されるまでにはまだ時間が必要だった。それでも KTB の掲出字テキストは「今昔文字鏡」[4]の支援を受けて、テキスト化がほぼ実現された。筆者がこの作業に関わり始めた 2011 年には、Unicode ver. 6.0 の CJK 統合漢字は既に Extension A-D まで拡張され、Extension A-B の実装も大きく進んできた。同年、KTB の掲出字テキストの UCS バージョン[6]が完成し、2016 年には Unicode の収録字数が 8 万字を超えるなか、ついに KTB の全文テキストが完成し、同年 9 月に全文テキスト公開が実現した[7]。

このような状況の中で『篆隷万象名義』の全文テキスト化に携わった筆者は、時代の波に乗っていたと言えるだろう。しかし、その当時は、このような大局的な視点を持っていたわけではなく、単に作業と研究に専念していた。表 1 で総括してみると、Unicode の CJK 統合漢字の収録字数とコンピュータ実装の進化を背景に、日本の DH 関連の多くのプロジェクトが拡大・発展してきたことが伺える。

2016 年 9 月に『篆隷万象名義』の全文プレーンテキストが公開された。この公開された『篆隷万象名義』全文テキストは、図 1 に示されているように、3 つの階層（文字列、単漢字、部品）での検索が可能である[8]。単漢字と部品に関する検索の研究を行っていた。

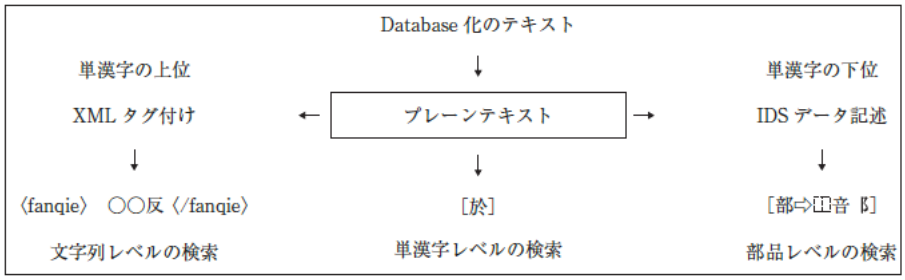

図 1 データベース化のテキストを利用する検索

博士課程を終えた後は、日本学術振興会特別研究員（PD）として京都大学人文科学研究所に、その後は、北海道大学文学研究院、関西大学アジア・オープン・リサーチセンター（KU-ORCAS）に所属した。そして、今年の 10 月 1 日、京都大学人文科学研究所附属人文情報学創新センターの助教に着任した。今後は引き続き古辞書研究と DH 関連のテーマで研究や業務を進める予定である。具体的には、日本古辞書の翻刻階層モデルの構築についての研究や、人文学のデジタルテクストに対する最新の構造化ガイドラインである TEI P5[9]を基にしたデータの構造化記述の方法に関する研究（図 1 の左部）を手掛ける。さらに、字形

のデータをより詳細に整備する作業も進行中である。

『篆隷万象名義』という研究テーマは、指導教員からの提案を受けて始めたが、古辞書研究を通じて、最先端の DH 研究と触れる機会を得た。さらに、古辞書研究と DH の進展は相互に影響を与え合っている。古辞書研究を深める中、漢字符号化の提案活動にも参加するようになり、漢字の符号化提案レビューや IRG の会議にも参加するようになった。北海道大学文学部の研究生として入学した際、『篆隷万象名義』の文字を大漢和辞典のデータと照らし合わせて符号化する作業を開始したことを思い出すと、感慨深いものがある。

これらの研究や活動は、ある意味、時代の流れや大きな波の中にあった。上記の表 1 で総括された内容をみると、Unicode の進展と日本の DH の発展との関連性が明らかである。IRG#61 に参加したことは、今もなお、その大きな動きや波の中にいると実感させるものがある。時にはその波が激しくなることもあるかもしれないが、ともに研究や活動を進めている仲間たちとともに、その困難を乗り越えることができると信じている。

[1] IRG#61: https://ceas.yale.edu/ideographic-research-group-61

[2] 池田証壽：篆隷万象名義データベースについて、國語學、178 号、pp.57-65（1994）。

[3] CJK Unified Ideographs: The name of the ideographs of Chinese origin that are included in Unicode, and divided among various blocks, specifically the URO and the multiple Extensions. Ken Lunde, *CJKV Information Processing 2nd Edition*, O'REILLY, p.761, (2008). CJK という表現は、China、Japan、Korea の頭文字を取ったものである

[4] Ikeda Shoju, Li Yuan, "Building a General Database System of Chinese Characters Dictionaries in Early Japan: Tenreibanshōmeigi in HDIC Project", Journal of the Graduate School of Letters , Vol.13, pp.49-64（2018）.

[5] 今昔文字鏡とは、コンピュータ上で漢字検索と印字を可能にするアプリケーションソフトウェアである。1997 年の発売当初は株式会社紀伊國屋書店が販売元であったが、その後版を重ね、現在は開発元である株式会社エーアイ・ネットにより販売されている。https://www.jepa.or.jp/ebookpedia/201704_3537/、日本電子出版協会、2023 年 10 月 30 日確認。

[6] UCS バージョンとは、UCS の範囲内での漢字符号化されたバージョンを意味する。UCS：Universal Character Set. Refers to Unicode and ISO 10646. Ken Lunde, *CJKV Information Processing 2nd Edition*, O'REILLY, p.791, (2008).

[7] 李媛・池田証壽：篆隷万象名義の全文テキストと公開システムについて、じんもんこん 2016 論文集、pp.95-102（2016）。

[8] 李媛：空海の字書 人文情報学から見た篆隷万象名義、北海道大学出版会（2023）。

[9] TEI P5：https://tei-c.org/guidelines/p5/

執筆者プロフィール

李媛（り・えん）京都大学人文科学研究所附属人文情報学創新センター助教。2017 年北海道大学大学院文学研究科博士後期課程修了、博士（文学）。専門分野は日本語学・人文情報学。著書に『空海の字書 人文情報学から見た篆隷万象名義』（北海道大学出版会、2023）がある。