「人文系データベース」における相互運用性をめぐる諸問題

永崎研宣
山口県立大学

下田正弘
東京大学大学院

近年、人文学を支援するためのデータベースは数多く Web に公開されている。さらに、それらの Web サービスを相互運用しようとする動きが徐々に出てきている。相互運用には、運用コストを削減しつつサービスを拡張できること、技術やその人文学への応用の仕方に関して情報交換ができること等といった様々なメリットがある。さらにそれは、人文学自体を見つめ直す機会になり得るなど、潜在的な可能性をも見出し得る。人文系データベース同士の相互運用にもいくつかの問題があり得るが、それでも、それは人文学とそのデジタル化にとっては大きな意義がある。

Interoperability between Databases for the Humanities

Kiyonori NAGASAKI
Yamaguchi Prefectural University

Masahiro SHIMODA
The University of Tokyo

Many database systems supporting the Humanities are published on the Web. Recently, interoperability is gradually enabling these databases to develop their services. Interoperability has some merits and possibilities, for example, reducing the cost of operation and maintenance, sharing information about the information technology and its application for the Humanities, or for reflection on the Humanities itself. Even if there are some problems in interoperability, it is significant for the Humanities and its digitization.

1．はじめに

本稿では、「人文学からの要求に潜在する可能性を顕在化させて人文学に提供し、それを受けてあらたに人文学から発信される課題を自身の発展に還元する」[1]という目的の下で構築され運用されるデータ、およびデータベースマネジメントシステムを「人文系データベース（以下、人文系 DB)」と総称する。

近年、環境問題への対応や情報インフラとしてのインターネットの幅広い普及等の様々な理由により、紙媒体による研究成果の公表が人文学研究における要請とは異なる文脈で制約を受けつつある中で、デジタル媒体を利用した成果公表はもはや避けて通れない道となりつつある。その

ような状況にあって、人文学の研究手法において紙媒体であるがゆえに暗黙的に前提とし得た部分をデジタル化に際して可能な限り分析しデジタル媒体において適切に再現していくことは、この過渡期にあって人文学に携わる者が次世代への責任を果たすために課せられた重要な使命の一つであり、人文系 DB の構築と運用においてはそのような視点が不可欠であるというのが本稿の基本的な立場である。

人文系 DB には様々なものがあるが、多かれ少なかれ、ある一定の方針に従って構築・運用されており、近年は Web で公開されて様々なサービスを提供しているものも数多く存在している。これらは基本的に、何らかの研究組織や補助金をバックグラウンドとしており、それらが依拠する研究分野からの何らかの要請を拠り所として構築されたものが大部分である。大規模なものはまだそれほど多くないが、小さなものも含めると無数に存在するといっても過言ではない[1]。このような状況を、サービスとしての「人文系 DB」を発展させていく[2]という面から観察してみるなら、ひとつの人文系 DB が際限なき拡張を展開していくよりもむしろ、そのような複数のサービスがそれぞれに連携して相互運用を行っていくことが現実的な選択肢の一つであるように見える。すなわち、データベースの規模が大きくなりすぎると、ひとつのプロジェクトだけで運用するのは技術・運用・権利・費用など様々な面で困難が予測されるからであり、その一方で、個々のプロジェクトの独立性を残したままにしておくことは、それぞれのプロジェクト及びそれが運用する人文系 DB がそのプロジェクトやその背景となる研究分野の方針に基づいて継続的に運用されることになり、運用方針の一貫性についてそれなりの期待が可能だからでもある。現在ではまだあまり事例としては多くないが、たとえば人間文化研究機構が 5 つの研究機関のデータベースを統合して 2008 年 3 月より公開している「研究資源共有化データベース」[4]は有力な先行事例の一つだろう。

ここでは、そのような、人文系 DB における相互運用性の問題に焦点を当て、主に大正新脩大藏經テキストデータベースにおける相互運用の事例を手掛かりとし、そこに見出し得る意義及び潜在的な可能性について検討を行いたい。

2．相互運用の意義

ここでは、大正新脩大藏經テキストデータベース(以下、SAT DB) [5][6]における相互運用の事例において、まず、相互運用の効果を通じてその意義について検討する。SAT DB は 2008 年 4 月に Web に公開された 600 万行、1 億 5 千万字以上の漢文仏教経典データを格納する典型的な人文系 DB であり、Linux、Apache、PHP、PostgreSQL、Ludia、Senna 等のフリーソフトを使ったオンライン全文検索サービスを提供しており、公開後の月間アクション数[3]は平均 10

---

1 たとえば、「国内人文系研究機関 WWW ページリスト」[2]に紹介されている Web サイトの数にその状況は端的にあらわれている。

2人文系 DB の発展については、「人文科学のためのデジタルアーカイブの弁証法的展開」という文脈で論じたので参照されたい[3]。

3 「アクション数」は、このデータベース上で利用者がなんらかの操作を行った回数を指している。

万を超えている。この人文系 DB の各種 Web インターフェイスは、AJAX[4]を全面的に活用し相互運用性を重視して構築・運用されている。ここで相互運用の対象となっているのは、現在のところ、漢字の異体字同時検索のための CHISE[7]、専門用語の参照のための DDB (Digital Dictionary of Buddhism)[8]、関連論文検索のための INBUDS[9]の３つのシステムである[5]。CHISE は、守岡知彦氏が開発し GNU GPL に基づいて公開されている「汎用文字符号に制約されない次世代文字処理環境」であり、現在のところ、SAT DB ではテキストの全文検索や DDB、INBUDS 等のデータの検索の際に、異体字を同時に検索するために CHISE の文字オントロジーデータを参照して検索を行うシステムを実装している。DDB は、A. Charles Muller 氏が 1995 年より Web に公開し続けている電子仏教用語辞典であり、世界中の協力者とのコラボレーションのもと、現在では 4 万以上のエントリを有する専門用語電子辞典となっている。SAT DB では、漢文テキストの読解を支援するために、AJAX を利用して、Web ページ上の本文テキスト部分をマウスで選択すると、選択されたテキスト中の語彙を DDB で検索して Web ページの右側にリスト表示させるという仕組みを実装した。また、同様にして、日本印度学仏教学会が構築・運用している専門分野特化型の論文書誌データベース INBUDS をも、連携して検索できるようにしている。

いずれのシステムも、それぞれに構築した開発者・組織が引き続き運用を行っており、データのアップデートも基本的には継続されている。したがって、SAT DB の運用者/組織は、相互運用のためのインターフェイスを一度開発しただけで、あとは、異体字検索、専門用語参照、関連論文検索に関しては、データのメンテナンスを行う必要がなく、その分の労力を SAT DB 自身の充実化や他のシステムとのさらなる連携によるサービスの充実化に割くことができるのである[6]。もちろん、連携システムをさらに増やしていく際にも、メンテナンスの労力があまり増えずに済むという点は大きなメリットである。

また一方で、SAT DB は、データベース上のデータを効率的に利用するためのいくつかの Web API を開発し、一部は完全に公開している。独自の、しかし仏教研究者にとっては比較的理解しやすい一定の書式、すなわち、紙媒体上でこれまで行ってきた参照方法に準拠した URI でリクエストをするとその箇所のテキストデータを HTML で整形して返すという仕組みである。この URI の記述方式は Web サイト上に公開されており、誰でもこの URI を利用して自らのシステムに SAT DB の一部を容易に組み込むことができるようになっている。筆者が知る限り、すでに二つの人文系 DB がこの API を利用して SAT DB のテキストデータを参照するようになっている。このように、API を公開することで、他の人文系 DB の運用者は、自らはほとんど運用コストをかけることなく SAT DB のデータを自由に利用できるのである。

以上のように、人文系 DB における相互運用は、より少ないコストでより便利なサービスを提

---

4 ここでは主に Yahoo! UI を利用している。

5 CHISE に関しては、ここで言う人文系 DB よりも大きな枠組みを目指して作られたものであり、その点には留意されたい。

6 仏教学分野の国際的なデータベース連携のワークフレームとしては Integrated Buddhist Archive が提唱されており [10]、西洋古典学においては「第四世代」としてワークフレームの構築が行われている[11]。

供するための手段として有力な選択肢となり得るものであることが、この事例からは明らかである。一方で、まだこの段階では顕在化していない、その潜在的な可能性について次に検討してみよう。

3．相互運用における潜在的な可能性

上述のように、ここでは、相互運用における潜在的な可能性について、運用面、各種技術やその応用の仕方、人文学研究における研究手法のとらえ方、といったいくつかの観点から検討したい。

3．1．運用面での可能性

すでに、前章において、運用面での相互運用の意義に関しては事例の範囲で検討した。しかしながら、本稿で採り上げている事例では未だ顕在化していないが、可能性として考えられるメリットやデメリットもある。

ここでは運用面を採り上げているのでその側面に限って考えるなら、たとえば、サービスやデータの永続性について個々のプロジェクトが責任を持つということがマイナスに働く場合もあり得るという点がデメリットとして考えられる。特に、プロジェクトが完全に終了してしまう場合、あるいは、単に公開をやめてしまう場合、公開方法を変更してしまう場合等が考えられる。こういった場合、個々のプロジェクトによるデータ公開そのものが危機に立たされることになってしまい、それぞれのプロジェクトによる独自運用を前提として構築された相互運用サービスにとっては、サービスそのものが提供できなくなってしまうという危険性がある。

その一方で、視点を変えるなら、いずれかの人文系 DB におけるデータ公開、あるいはプロジェクトそのものの継続的運用が危機に瀕するというようなケースにおいては、もし相互運用がすでに実施されていたなら、連携先のプロジェクトの運用状況を日頃から相互に把握しやすいため、必要に応じて連携先プロジェクトを何らかの形で支援したり、最終的には運用そのものを肩代わりしたりするといったことも可能であり、このことは相互運用においてあり得るメリットの一つだろう。すなわち、近年特に問題とされているデータの持続可能性の問題[12]を解決しようとする際にも人文系 DB 同士の相互運用は寄与し得るのである。

3．2．各種技術やその応用手法等

相互運用を開始するためのプロセスにおいては、技術的な面、とりわけそれぞれの研究分野への応用の仕方に関して様々な情報交換が必要となる。そのことは、それぞれの人文系 DB が採用する技術やその応用手法をお互いに学びあう機会ともなり得る。あるいはさらに、人文系 DB の構築・運用のためのワークフローについての情報交換も行えることになる。常にうまくいくとは

限らないが、相互運用に取り組むそれぞれの分野が双方の持つ特長を互いに学びあうことも可能であろう。また、一方が技術的、ないしワークフローのような面で問題を抱えてしまっている場合には、それをフォローすることが期待でき、さらに、当事者が気付いていないような問題についての指摘を行う機会ともなり得る。いずれにしても、技術に関連するような側面においても、相互運用がもたらし得るメリットは少なくない。

3．3．人文学の研究分野における研究手法のとらえ方

本稿で採り上げた事例ではまだあまり問題になっていないが、本来、人文系DBは、サービスの対象とする学問分野の方法論に依拠して構築されるものであり、同じ資料を使って人文系DBを構築したとしても、依拠した方法論が異なれば、構築手法からデータの構造、ユーザインターフェイスに至るまで、まったく異なったものになってしまう可能性がある。たとえば、SAT DBにおいて現時点でデジタル化されているのは「紙媒体に活版印刷された校訂テキストとしての線形のテキスト本文とそのヴァリアントを主とする脚注、そして、適宜挿入された図像」であり、その線形テキストの位置情報はテキスト番号、巻番号、ページ・段・行によって記述されている。これは、元になる資料の性質と、その資料を扱う研究分野、すなわち、SAT DBの母体である仏教学分野の研究手法に依拠しており、そこでは問題なく利用可能である。しかしながら一方、たとえば、正倉院文書データベースSOMODA[13]を見てみるなら、元になる資料においてテキストは線形ではあるものの複数テキストが複雑に入り乱れており、未だ線形テキストが完全に取り出せているとは言えないとのことである。ここにおいてデジタル化の対象となり得るのは、現時点で取り出し得た線形テキストよりも、むしろ、入り乱れた資料そのものであり、したがって、検討すべきなのは、その資料の入り乱れ方をいかにして可能な限り忠実に、利用する研究者が共有できる形でデジタル化するか、という点であり、そして、それらの入り乱れた状態において研究者が脈絡をつけやすくするための様々なレベルでのインターフェイスを用意することが肝要だろう。SOMODAが目指しているのはまさにそういった方向性であると思われる[14]が、そのような形でのデジタル化が進められている人文系DBは、SAT DB、すなわち、すでに、元になる資料が写本ではなく活版印刷された校訂テキストでありその扱い方についても研究分野において手法が確立しているような資料をデジタル化した人文系DBとは大きく異なっている。人文系DBとは、このように、それぞれの資料の在り方、そしてそれを対象とするそれぞれの研究手法を如実に体現した個別的なものとならざるを得ないのである。[7]

SAT DBが現在相互運用を行っている対象は、辞書や字体情報、論文書誌情報であり、比較的相互運用しやすい種類のものだが、たとえばSOMODAのようにまったく異なる人文系DBとの領域横断的な相互運用を行なおうとするなら、そこでは運用面の問題や技術面での交流だけでなく、資料の在り方やそれに基づく研究手法そのものについて相互に深く理解することが必要となるだろう。そもそも連携することの意義がどこにあるのかという検討も必要だが、それは一方

7 この点については研究分野の意味論の構築という視点から論じている興味深い論考がある[15]。

で、個々の人文系 DB が他者を通して様々な次元において自らを見つめ直す契機ともなり得る。その次元には、技術的な新たな見通しや研究手法への応用の仕方から人文系 DB 構築・運用のためのワークフローの様態に至るまで、様々なものがあり得る。さらにまた、それらが依拠する個々の研究分野の研究手法をも参照することになる。とりわけ、異なる研究手法に依拠する人文系 DB 同士の相互運用という局面においては、それぞれの研究手法にとって「より良いサービス」とは何か、ということが改めて問われることになる。研究手法が異なっている以上、「良さ」の方向性にずれが生じることは十分にあり得る事態である。この場合には、ひとつの「良さ」のみを追求するだけでは不十分であり、それぞれの「良さ」をも反映させ得る形での相互運用が必要であり、また、同時に、双方における「良さ」を再考するきっかけにもなり得るだろう。そのように、人文系 DB 同士を連携させることを目指す営みは、必然的に、デジタル化の範疇のみにおさまらず、個々の研究分野そのものが領域を超えて学際的にコラボレーションをするという形にならざるを得ず、さらにそれは、「利用者への具体的なサービスの提供」という形で成果が明示されることになるだろう。

このようにして、それぞれの人文系 DB が依拠する研究手法に即しつつ、個々に丁寧に相互運用を展開していくことができたとしたら、それらを架橋していくような形で「より広い人文系 DB」の枠組みを目指すことが可能かもしれない。そして、それを通じて、人文学におけるさらに広い研究手法についての検討が可能となるかもしれない。実装面においては、暫定的な目標とし得る入れ物として、我々の前にはすでに World Wide Web が用意されており、ここに適切な形で導入することができたなら、より高度な利用と適切な情報発信が実現していくことだろう。とはいえ、そこに至るための道具立てを充実させるためには、各分野での一層の努力が必要だろう。

4．終りに

ここまで見てきたように、人文系 DB の相互運用は、個々の人文系 DB の機能強化と利用者へのサービスの向上をもたらすことができる。また、構築・公開までは熱心だがその後は放置されてしまいがちであり有効性の検証が行われる機会も少ない人文系 DB にとっては、相互運用は継続的・発展的に運用するための動機づけとして有効であり、比較的近い問題意識を持ったグループによる実効性の高い検証が行われる機会ともなり得る。そのように、人文系 DB の相互運用とは、単に提供するサービスを充実化するだけにとどまらず、様々な点で人文系 DB にとっての意義が大きいということは明らかである。さらに、それのみならず、人文学の方法論そのものにとっても有益なものをもたらし得るいくつかの潜在的な可能性がある。この可能性については、未だ明確な答えを提示できる段階にはないが、今後も課題としていきたい。

謝辞

本稿執筆にあたっては、東京大学大学院人文社会系研究科・文学部次世代人文学開発センター・萌芽部門・データベース拠点ワークショップシリーズ（第Ⅰ期）における 2 度にわたるワークショップ「人文学の研究方法と「人文系データベース」の設計思想」においてそれぞれの提題者を務めてくださった守岡知彦氏、後藤真氏との議論から多くの示唆を受けたことを感謝とともに記しておく。また、同センターの研究員である白須裕之氏からも多くの助言をいただいた。A. Charles Muller 氏には DDB との連携にあたり貴重なデータを提供していただいた。なお、本研究の一部は、文部科学省科学研究費補助金萌芽研究「次世代新大蔵経編纂スキームの構築（研究代表者：下田 正弘）」（課題番号：20652005）の一部として遂行された。

参考文献

[1] 東京大学大学院人文社会系研究科・文学部　次世代人文学開発センター・萌芽部門・データベース拠点ワークショップシリーズ（第Ⅰ期）「人文学の研究方法と「人文系データベース」の設計思想」
http://www.l.u-tokyo.ac.jp/cgi-bin/report.cgi?mode=2&id=163 （2008 年 11 月 10 日参照).
http://www.l.u-tokyo.ac.jp/cgi-bin/report.cgi?mode=2&id=168 （2008 年 11 月 10 日参照).

[2] 後藤斉 [1995] 「国内人文系研究機関 WWW ページリスト」, http://www.sal.tohoku.ac.jp/~gothit/zinbun.html （2008 年 11 月 10 日参照).

[3] 永崎研宣[2005]「デジタルアーカイブの弁証法」『情報処理学会研究報告』CH-68(2005 年 10 月), pp. 17-24.

[4] 人間文化研究機構 「研究資源共有化データベース」 http://www.nihu.jp/kyoyuka/tougou/ （2008 年 11 月 10 日参照).

[5] 大藏經テキストデータベース研究会（SAT）「大正新脩大藏經テキストデータベース」http://21dzk.l.u-tokyo.ac.jp/SAT/ （2008 年 11 月 10 日参照).

[6] Nagasaki, K. and Shimoda, M. [2008]. Outline of the Activities of the SAT Project, Joint International Conference on Digital Buddhist Studies, at Dharma Drum Buddhist College, February 2008: 22-23.

[7] 守岡知彦[2006]「文字オントロジーに基づく文字処理について」『情報処理学会研究報告』CH-72(2006 年 10 月), pp. 25-32.

[8] Muller, A. Charles. [2008]. The Digital Dictionary of Buddhism [DDB]: Present Status and Future Developments, The Ninth Annual Symposium for Scholars Resident in Japan, March 2008.
http://www.acmuller.net/articles/ddb-nichibunken-200803.html (2008 年 11 月 10 日参照).

[9] 日本印度学仏教学会データベースセンター（INBUDS）Home Page http://www.inbuds.net/ （2008 年 11 月 10 日参照).

[10] Muller, A. Charles. [2008]. EBTI After 15 and CBETA after 10 Years: Joint International

Conference on Digital Buddhist Studies (February 15-17, 2008): Chair's Report, http://buddhism-dict.net/ebti/ ebti2008report.html (2008年11月10日参照).

[11] Crane, G. [2008]. Fourth Generation Collections: TEI, FRBR, and Canonical Text Services, TEI Member's Meeting 2008, Nov 2008, http://www.cch.kcl.ac.uk/cocoon/tei2008/programme/abstracts/abstract-160.html (2008年11月10日参照).

[12] Rehm, G. and Witt, A. [2008]. Aspects of Sustainability in Digital Humanities, Digital Humanities 2008 , June 2008: 21-29.

[13] 正倉院文書データベース(SOMODA) http://somoda.media.osaka-cu.ac.jp/ (2008年11月10日参照).

[14] 後藤真[2008]「正倉院文書データベースと「復原」」『アジア遊学 No.113』(2008年8月), pp. 52-59.

[15] 白須裕之[2008]「人文系データベースを構築するとはどういうことか？」『漢字文献情報処理研究』(2008年10月), pp. 11-19.