大蔵経における多言語対訳コーパスの構築

永崎 研宣*1　　白須 裕之*2　　下田 正弘*2

*1 人文情報学研究所　　*2 東京大学大学院人文社会系研究科

仏教学分野では、中国語・チベット語の大蔵経をはじめとして歴史的に様々な言語に翻訳されたテクストが存在し、その多くがデジタル化資料として公開されつつあり、目下、これらの利便性をどのようにして高めていくかということが一つの課題となっている。その中で、それぞれのデジタル化資料を相互運用していくという方向性も模索されつつある。ここでは、異なる言語間の資料の関連づけに焦点をあて、多言語対訳コーパス構築の手法について検討した上、今回新たに構築しつつある関連づけのためのシステムについて紹介する。

A Multilingual Parallel Corpus System for Daizokyo

Kiyonori Nagasaki*1　　Hiroyuki Shirasu*2　　Masahiro Shimoda*2

*1 International Institute for Digital Humanities

*2 Graduate School of Humanities and Sociology The University of Tokyo

As more and more digital materials are developed by various projects in the field of Buddhist Studies and distributed onto the Web, it becomes increasingly necessary to create a relationship between such projects and their databases. It is especially useful for scholars of Buddhism to correlate the various language versions of a single text with each other. In this paper, the current situation of the method for making this correlation will be discussed. This will be followed by a discussion of how to create a support system to that will allow follow direct manual scholarly input in a collaborative manner.

1. はじめに

大蔵経とは仏典の集成を指す言葉である。中国語やチベット語等の文化圏においてそれぞれの言語に翻訳された仏典をまとめて扱うために用いられるようになったものであり、その時点で重要と思われる仏典が集められ、リストされている。国家事業として行われることが多く、後には木版によって印刷されるようになり、仏教の普及と伝承に大きな役割を果たした。当初、仏典の伝承にはサンスクリット語やパーリ語等が用いられたとされているが、これらの言語によるテクストは必ずしも残存しておらず、また、残存していたとしても、仏教の伝播の過程で時代や地域によって生じたであろう様々な派生形まで網羅できるとは限らない。こうしたことから、中国語やチベット語に翻訳されて現存するテクストは、サンスクリット語・パーリ語等によるテクストが残存していない場合や断片的にしか伝承されていない場合には、テクストそのものの存在と内実を明らかにするものとして、あるいは、そうしたものが現存する場合にも、翻訳時点での時代的地域的なテクストの内実を少なからず反映させているものとして、仏教研究においては重要な価値を持っている。一方で、そうした伝承の中での翻訳テクストのみならず、いわゆる近代仏教学の成立以降には、日本語訳や英語訳をはじめとする様々な現代語訳が世界中で出版されている。しかしながら、こうした翻訳仏典テクストは、様々な言語系統や文化圏に属するものであり、テクスト間の比較対照は必ずしも容易なものではない。すでに多くの仏典がデジタル化されつつあるという現状において、それらのテクストをどのようにして有益な形で比較対照していくかということに関しては、今後の大きな課題となっている [1]。

以上のような状況に鑑み、本稿では、それらの電子テクストから多言語対訳コーパスを構築する

[1] システム構築の目的は異なるが、すでにサンスクリット語とチベット語の仏典の構文対照電子辞書を目指すプロジェクトが進められている[1]。

ための構想を検討するとともに、現時点での実装系について報告する。なお、大蔵経で使用されている言語を対象とした自然言語処理、特に形態素解析のシステムは、大蔵経の地理的・時代的な広がりを前提としたとき、現段階では、十分に実用的なものが提供されているとは言い難い。ここでは、そのような制約の下で多言語対訳コーパスを構築するために、現在利用できる方法、技術を検討し、その方式、構築方針について議論する。

2. 対訳コーパスとその構築

現在では、文献[2]において提示されているように、言語学におけるコーパスの果す役割は非常に大きなものになっている。特に、対訳コーパスは様々な言語情報、語彙知識等を抽出するための基盤ともなっている。例えば、対訳テクストから対訳辞書の自動生成、翻訳支援、自動翻訳等への応用である。その基礎となるのが並行テクストの存在である。並行テクストは対訳された文献の文書構造にあわせて、対訳関係(どのテクスト断片が対応する翻訳であるかという情報) を合わせ持つテクストである。この並行テクストの構築にあたっては、自然言語処理や言語的単位に対する統計処理が前提とされている。本節では、対訳コーパスの構築のための要素技術について述べ、大蔵経の対訳コーパスの構築のために必要な項目を検討する。

2.1. 対訳関係の推定アルゴリズム

並行テクストの対訳関係を推定するアルゴリズム(テクストアライメント) は、対象となる文書の性質によって様々なものが提唱されている[3]。文レベルの対応については、文に含まれる文字数等の内部情報を使用するアルゴリズムが出発点であり、その後、訳語情報を考慮したアルゴリズム等が提唱されてきている。

しかしながら、大蔵経において使用されている各種言語においては、地域的、時代的な制約のため、自然言語解析による統語情報を十分な形で使用することが望めない場合がある。従って、文やパラグラフの対訳関係を対象とする場合、文字数等からの統計情報を利用するアルゴリズムと、人手による修正などの両面を両立させる方式を検討する必要がある。また、対訳辞書によるアンカーを利用することも今後の課題である。

2.2. 対訳テクストのモデルに対する要求

すでにこれまで、文書データベース（以下、文書 DB）における要素間の関連情報データベース（以下、関連情報 DB）に関する研究[4][5]において提出してきた要求条件、及び文献[6]において指摘した問題を再整理した上で、対訳テクストをも含む関連情報 DB のモデルを設計する上での要求条件の主な事項を以下に挙げる。

1. RESTful なシステム[7]
2. より汎用的な様々な文書 DB との動的な関連づけ
3. 文書 DB 内での動的な関連づけ
4. 知識ベース的なものとの動的な関連づけ
5. 関連情報データに付随する詳細情報
6. 個別の関連情報データに関する信頼性の確保

なお、RESTful なシステムは、汎用的な相互運用性の確保を志向するために採用している[8]。

3. アライメント・ガイドシステム

3.1. システムの概要

現状では、機械処理による手法では、並行テクストのアライメントの部分的な実現が可能であるに過ぎない。ここでは機械的なテクストアライメントを補助する目的で、人手によるテクストアライメントを実現するためのシステムを提案する。今回のシステムでは、現代に和訳された仏典と漢訳され伝承されてきた仏典、すなわち現代的な解釈を反映した文章と仏教の伝統の中で伝えられてきた漢文との対訳コーパスを構築することを当面の目標とし、そのために、両文書 DB の関連情報を記述することを目指す。漢文のデジタル化においてしばしば問題となるのはテクストを指示する際の単位の設定の困難さだが、今回の場合には、現代日本語との対応ということになるため、現代日本語における単位、すなわち、段落、文章、あるいは術語といったものに対して、対応する漢文の情報を関連づけるという形でこの問題をある程度回避できることになる。また、すでにここで取り扱っている SAT Web DB と動的に連携している電子仏教辞典[9]を利用することで、ある程度作業を支援することが可能となる。これにより、構築作業者は、和訳において任意の箇所を指示し、それに対応する漢文情報を対応付けし、それに関する補足的な情報を追記することになる。この際、指示された箇所は URL によって記述され、蓄積されることになる。

3.2. 関連情報の内容

本稿 2.2. に述べた経緯により、関連情報の内容については以下の枠組みによって記述した。

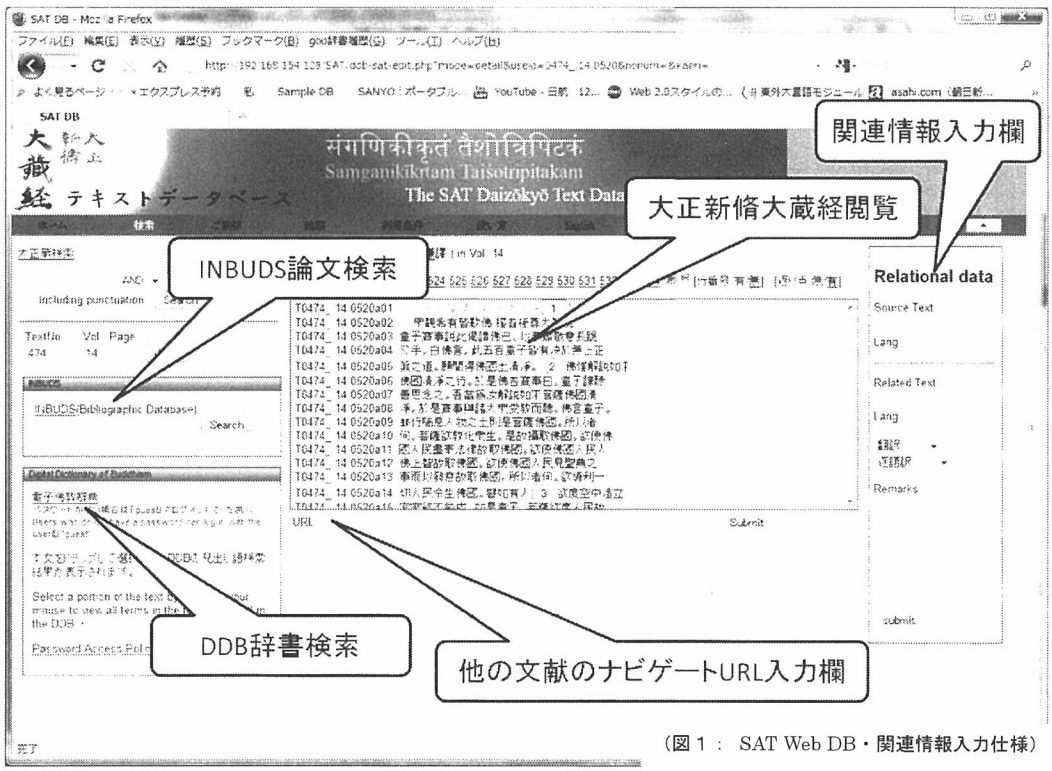
A 軸：（オーソライズ情報）
A-1 研究者による入力（データ著者）
— 著者名、時間、内容、備考
A-2 オーソライズ組織によるチェック
— チェック者名、時間、内容、備考
A-3 オーソライズ組織による公開
— 公開者名、時間、内容、備考

B 軸：要素間の関連の仕方についての情報
B-1 翻訳（逐語訳、意訳）
B-1-1 各要素の言語情報(基本は各文献の通りだが上書きの場合に記述)
B-2 引用（ほぼ同一、少し異なる、要約）
B-3 異読（少し異なる、大きく異なる、不足、余分）

B-4 注釈（入力者の任意）
B-5 直接対応（片方が知識ベースのようなものの場合）

C 軸：データの重み付け情報
C-1 ユーザによる重み付け操作
C-2 データの利用状況による重み付け

ここで挙げているのは多言語対訳コーパスをも対象として含む関連情報 DB であり、ここでは、以上の関連情報に基づき、主に仏教学関連の文書 DB（文書画像データベースを含む）を対象として記述・蓄積していくことになるが、さらに、これまでの成果に加えて文献[5]において指摘した情報の妥当性を確保すべく、A 軸を再構成しつつ C 軸を新たに設定している。A 軸において「入力」→「チェック」→「公開」という三段階を経ているのは、2008 年 11 月より開始された Web コラボレーションである日本インド学仏教学会の INBUDS[10]のワークフロー及びコラボレーションシステムの運用による成果を反映している。これによって、関連情報として蓄積される情報の信頼性の確保がある程度可能となる。A 軸はデータの作成者側によって提供される信頼性であり、C 軸は利用者側から提示される信頼性（もしくは評価）となる。この関連情報 DB では、このような形で、どの情報がどの程度有益かということを利用者に把握しやすくするとともに、複数の異なる解釈を共存させるための枠組みを提供することを志向している。B 軸に関しては、やや主観の余地が大きく、これをどう考えるかは運用段階での課題としたい。



（図 1 ： SAT Web DB・関連情報入力仕様）

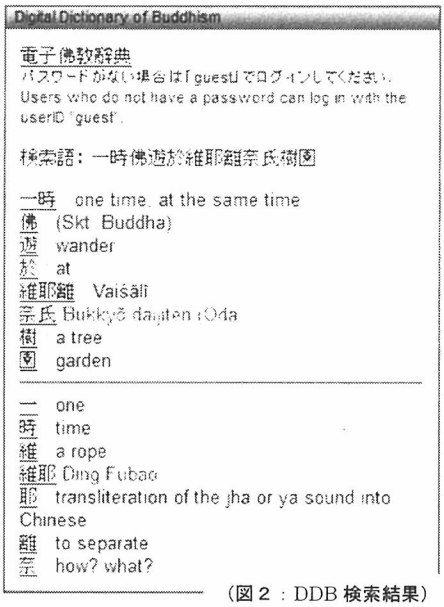
3.3. システムの現状

今回のシステムは、文献[4][5]において提示したシステムを発展させるとともに、人手による作業を効率化するための様々な工夫を行っている段階である。

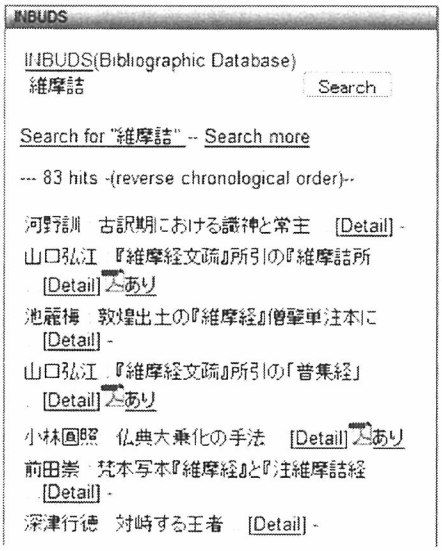
まず、SAT 大蔵経テキストデータベース研究会によって公開されている SAT Web DB[11]を全面的に利用することによって、ユーザにとっての利便性を可能な限りそのまま引き継げるようにしている。SAT Web DB は、大正新脩大蔵経のテキストデータベースであり、利便性向上のために、電子仏教辞典（以下、DDB と呼ぶ）や日本インド学仏教学会論文データベース（以下、INBUDS と呼ぶ）等をシームレスに活用できるインターフェイスを備えている（図 1）。なお、DDB を利用す

る場合には、大蔵経の本文をドラッグすることで辞書検索を行い、当該ウインドウに辞書検索結果が表示されるようになっており、現バージョンでは、選択テクストの冒頭からの辞書見出し語との最長一致で術語ごとに分割しそれぞれの英語の意味を表示するようになっている（図２）。また、INBUDS に関しては、本文をドラッグすると検索窓にドラッグしたテクストが入力され、「Search」をクリックするとその単語をキーワードとして持つインド学仏教学分野の関連論文が表示され、さらに、CiNii Web API[12]を自動的に参照して論文 PDF ファイルへの直接リンクを用意するようになっている。（図３）すなわち、利用者は、仏典を読みながら、辞書を引き、関連する論文を参照したりすることが容易に可能となっているのである[13]。このような機能は、対訳コーパス構築においては極めて有益であることから、これらの機能をそのまま利用できるインターフェイスとした。

なお、図１中、通常の SAT Web DB と異なっているのは、「他の文献のナビゲート URL 入力欄と「関連情報入力欄」が加えられている点である。



（図２：DDB 検索結果）



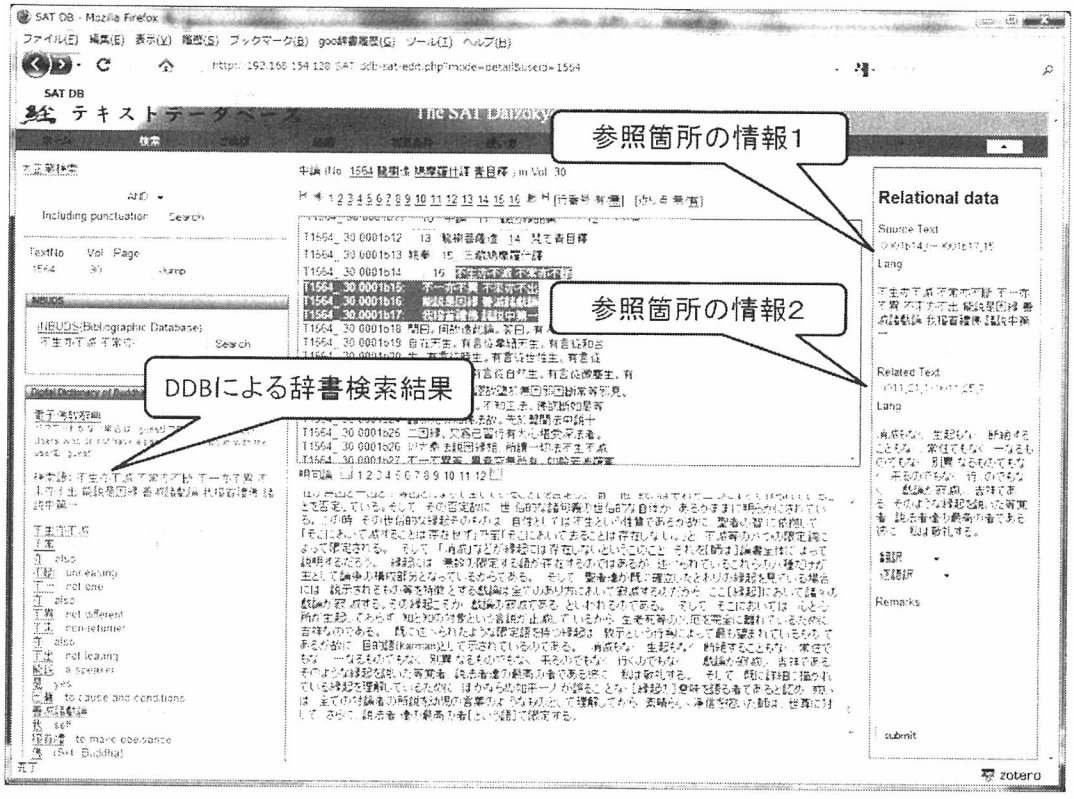
（図３：INBUDS 検索結果）

3.4. 関連づけ対象テクスト DB の要件

SAT Web DBは、AJAXを活用することで、テクストのドラッグをトリガーとして動的に様々なことができるように設計されており、関連づけ作業においてもこれを生かすことで効率的な作業が実現可能である [2]。しかし、関連づけ作業を効率化するためには、関連づけ対象となるテクストのデータベース（以下、関連づけ対象テクストDBと呼ぶ）においてもある一定のフォーマットが必要となる。このことは、今後関連づけ対象となるデジタル化資料を増やしていくにあたっても重要となる。ここではあくまでも仮の手法として、関連づけ対象テクストDBに対して、①ページ番号等による本文参照機能と②本文表示機能の二つを要求するという形式をとった。すなわち、図１における「他の文献のナビゲートURL入力欄」に①の機能をもったコンテンツが表示されるURLを入力し、「Submit」をクリックすると、その個所にナビゲートのためのコンテンツが表示され、さらに、そのコンテンツ中のページ番号等をクリックすると、下のボックスに、それに対応する関連づけ対象テクストDBの本文が表示されるというものである。なお、この一連の作業におけるコンテンツの遷移はすべてAJAXによって実装されており、Webページそのものの遷移は行われないようになっている。これを実現するためには、SAT Web DBと関連づけ対象テクストDBとの間でどのように役割分担をするかという問題があるが、今回は仮に、関連づけ対象テクストDBの方では、

[2] なお、AJAX ライブラリとしては、当初は Yahoo! UI [14] を利用していたが、その後、動作の軽量化のため jQuery[15]への移行を進めており、80%程の移行が完了している。

たとえば 75 ページの場合は 75 としておき、この数字をクリックした場合、SAT Web DB側に用意されたJava Scriptが関連づけ対象テクストDBに対して本文を改めて問い合わせるという形にした。この手法は、基本的には、XML等で作成されたファイルをはじめ、どんな様式のデータであっても、これにあわせるためのラッパーを作成することによってここでの要件を満たすことは可能だが、この件については、あくまでも過渡的なものであり、今後、より良い手法を検討したい。



（図４：関連情報の入力）

3.5. 関連情報の入力

関連情報として蓄積される情報としてもっとも重要となるのは、関連情報の元となる個々の要素の情報をいかにして簡単かつ正確に入力するか、である。まず、SAT Web DB においてデジタル化・公開されている大正新脩大蔵経の場合には、研究者によるテクストの参照方法として、ページ・段・行という形式（たとえば、1 ページ b 段 14 行目、等）がほぼ標準化されており、デジタル化資料として扱う際にも、この形式をそのまま援用することが可能である（たとえば、001b14 等 。また、機能面では、上述のように、SAT Web DB では、すでにマウスによる本文ドラッグをトリガーとした機能が用意されており、これに機能を加える形で参照箇所を表示・入力することが可能である。したがって、SAT Web DB に関しては、ページ・段・行、さらに、これに加えて、行頭からの字数を位置情報として、始点と終点を記述することでひとつの要素として扱うこととした。（たとえば、001b14_0-001b17_15 等）

一方、関連づけ対象テクスト DB の場合には、必ずしもこの手法が有効とは限らないが、紙媒体で流通している資料の場合にはページ・行による参照は比較的使いやすいため、ここでも当面は、ページ・行、さらに、これに加えて文字数という形で参照箇所を記述することとする。（たとえば、0011_21_1-0011_25_7 等）また、こちらに関しても、当該個所をドラッグして選択することで参照箇所の情報が入力される機能を用意した。

その他の関連情報については、すでに「3.2 関連情報の内容」において述べたとおりである。それ

らのうちの一部は、自動的に生成・入力することが可能である。とりわけ、A 軸の情報は、認証情報や入力時間など、自動化することで敢えて改めて入力する必要がなくなるものがほとんどである。また、C 軸に関しては、閲覧者側が入力するものと、利用状況からサーバ上で動的に生成するものとなり、関連情報入力時に入力するデータではないため、今回は割愛する。B 軸に関しては、入力手法としては単なるセレクトメニューとして用意されるに過ぎないが、入力作業者の判断を要するという点で難しくかつ重要である。より大規模な運用段階に入った後にも、選択項目の妥当性について引き続き検討が必要であり、また、機械処理との併用が持つ可能性についても検討していきたい。

4. おわりに

本稿では、対訳コーパス構築のためにデータを蓄積するシステムを紹介した。漢訳大蔵経においては、段落、文章といったテクストの基本的な枠組みが必ずしも明確ではないことが、機械処理を行おうとする際に様々な局面で問題となるが、これに対して、日本語や英語などの現代語訳との対訳コーパスが提供されることで、翻訳者による一つの解釈ではあるものの、ある一つのテクストの構造を前提として機械処理を行うといったことも可能となるだろう。

すでに述べたように、AJAX の発展によって、Web でのコラボレーションにおけるインターフェイスの飛躍的な改善が比較的容易に可能となったことは特筆に値するものであり、今後もより一層期待できるだろう。

しかし一方で、上述のように、参照箇所をどのように記述するかということや、実際にデータをどのように受け渡し、どのように効率化するかということについては、複数のデジタル化資料公開システム間でかなりの議論が必要となるだろう。近年の Web API 公開の普及はこの問題に解決の糸口を与えてくれることだろう。

また、関連情報の内容における、A 軸、C 軸という枠組みに関しては、関連情報 DB の信頼性の確保を目指すものではあるものの、それのみにとどまらず、デジタルメディア時代に適用し得る、広汎な意味でのアカデミックな評価システムの基盤へとつながっていく可能性も考慮に入れてさらに検討を進めていきたい。

以上のように、コーパス構築には多くの検討事項を克服しなければならない。しかし、実現されれば、仏教研究、言語研究に大いなる貢献ができると確信するものである。本稿はその長い道程への一歩である。

参考文献

[1] Tibetan-Sanskrit 構文対照電子辞書プロジェクト eDic, http://suzuki.ypu.jp/edic/ 2009 年 11 月 18 日参照.

[2] D. Biber, S. Conrad, R. Reppen: Corpus linguistics- investigating language structure and use, Cambridge University Press, 1998. (邦訳齊藤俊雄他共訳, コーパス言語学- 言語構造と用法の研究, 南雲堂, 2003.)

[3] J. Véronis, ed.: Parallel Text Processing — Alignment and Use of Translation Corpora, Kluwer Academic Publisher, 2000.

[4] 永崎研宣: 要素間の関連情報を基盤とする仏教文献デジタル・アーカイブの可能性, 情処研報 CH75, pp. 31-38, 2007.

[5] Kiyonori Nagasaki, A Collaboration System for the Philology of the Buddhist Study, Digital Humanities 2008, pp. 262-263, 2008.

[6] 永崎研宣, 人文科学のためのデジタル・アーカイブにおけるステイクホルダー —仏教文献デジタル・アーカイブを手掛かりとして—, 人文科学とコンピュータシンポジウム論文集, pp. 347-354, 2007.

[7] T.R. Fielding, Architectural Systems and the Design of Network-based Software Architectures, Dissertation, Information and Computer Science, University of California, Irvine, 2000.

[8]守岡知彦, データを生み出すデータのために, 人文科学とコンピュータシンポジウム論文集, pp. 13-18, 2008.

[9] 電子仏教辞典 Digital Dictionary of Buddhism, http://www.buddhism-dict.net/ddb/ 2009 年 11 月 18 日参照.

[10] インド学仏教学論文データベース INBUDS, http://www.inbuds.net/ 2009 年 11 月 18 日参照.

[11] SAT 大蔵経テキストデータベース SAT Web DB, http://21dzk.l.u-tokyo.ac.jp/SAT/ 2009 年 11 月 18 日参照.

[12] CiNii 外部インターフェイス提供について, http://ci.nii.ac.jp/info/ja/if_link_receive.html 2009 年 11 月 18 日参照.

[13] Kiyonori Nagasaki, A. Charles Muller, Masahiro Shimoda, Aspects of the Interoperability in the Digital Humanities, Digital Humanities 2009, pp. 375-377, 2009.

[14] Yahoo UI, http://developer.yahoo.com/yui/ 2009 年 11 月 18 日参照.

[15] jQuery, http://jquery.com/ 2009 年 11 月 18 日参照.