Subversionを用いた仏典テキスト校訂支援システム

丁 敏[1] 村川 猛彦[2] 福岡 整[2] 中川 優[2]
[1]和歌山大学 大学院システム工学研究科 [2]和歌山大学 システム工学部

奈良平安古写経撮影画像と大正新脩大蔵経テキストファイルを対照表示させ，経典画像と同一内容のテキストファイルを効率よく作成することを支援するWebアプリケーションを構築した．校訂管理のため，バージョン管理システムSubversionを導入し，WebサーバがSubversionのリポジトリと作業コピーを持つことで，各利用者はブラウザのみで利用できるようにした．テキスト校訂に配慮した，仏典全文検索システムの改良も試みている．

Support System for Revising Buddhist Canons using Subversion

Min Ding[1] Takehiko Murakawa[2] Hitoshi Fukuoka[2] Masaru Nakagawa[2]
[1]Graduate School of Systems Engineering, Wakayama University
[2]Faculty of Systems Engineering, Wakayama University

We develop a Web service for helping one to make the document which is the same in content as the shot image of a Buddhist sutra transcribed in Nara and Heian Periods, where the document derives from the text dataset of Taisho Tripitaka. Since Subversion is introduced for revision management and the server holds the repository and the working copies, the service is available only with Web browsers. Moreover we attempt to improve our full-text search system in order to deal with the variation of text files of the same copytexts.

1. はじめに

筆者らは，国内の古写経を対象として，その撮影画像を既存のテキスト情報と対応付けるための機械的・半機械的手法の確立を目指している．画像は視認性に優れているが，そのままでは検索ができない．大正新脩大蔵経テキストデータを用いた全文検索が有効であるが，画像とテキストの小さな不一致により，テキストファイルをうまく取得できないことがある．

これまで，経典撮影画像と大正新脩大蔵経テキストファイルを対照表示させ，経典画像と同一内容のテキストファイルを効率よく作成することを支援するWebアプリケーションを構築してきたので本稿で述べる．筆者らが構築してきた仏典全文検索システムの改善についても報告する．

2. 準備

2.1 対象とする漢訳仏典データ

本研究で対象とする，漢訳仏典の電子データは大きく分けて2種類があり，その流通経路および電子化方法が異なっている．それぞれについて紹介する．

中国で漢訳された仏典が，中国への修行僧による書写もしくは輸入により日本に入ると，その経典がさらに書写され国内で流布していった．仏教学，国文学などにおいて特に重要なのは，奈良・平安・鎌倉時代に書写されたものであり，「奈良平安古写経」と呼ばれる．近年，その存否状況の調査と，デジタルカメラを用いた経典撮影（電子画像化）が座主の協力のもと行われ[1][2]，筆者らはその画像処理について検討し，試作システムを構築してきた[3]．金剛寺一切経の経典撮影画像は3千巻を超え，その中のいくつかの経典はテキスト化が試みられているが，網羅的なテキスト作成には至っていない．

もう一つの漢訳仏典の流通の経路・媒体は，刊本一切経である．刊行時期は上述の古写経より少し遅れて10世紀以降とされている．20世紀に編集され，仏教学において漢訳仏典を引用する際に広く用いられる「大正新脩大蔵経」は，高麗版を中心とした刊本一切経を底本としている．さらに大正新脩大蔵経については，日本の大蔵経テキストデータベース研究会(SAT)，および中華電子仏典協会(CBETA)により，そのテキスト化が行われ，約9千巻にも及ぶそのテキストデータは無料で入手可能である．

奈良平安古写経と大正新脩大蔵経は，漢訳仏典という共通点を持ちつつも，書写時や木版に至るまでに記載ミスや意図的な変更が行われ，大小の相違点（異同）が存在する．そこで，大正新脩大蔵経の経典テキストデータを，対応する古写経経典データ（画像）に合わせるよう修正すれば，その修正テキストデータを用いて経典画像の検索が可能となるだけでなく，修正前後のテキストデータの異同を計算機により効率よく発見し，あるいは管理することにつながる

と期待される．

2.2 漢訳仏典全文検索システム

経典画像から対応するテキストファイルの特定を支援するため，筆者らはこれまで，全文検索システムを構築した[4]．CBETA の大正新脩大蔵経より抽出したテキスト情報をデータベースに登録し，検索エンジン Senna の近傍検索を用いて，ワイルドカード検索や複数行検索ができるようにした．インタフェースを図 1 に示す．16 枚の経典画像に対応するテキスト文字列をもとに，そこに現れる部分文字列を検索することで元のテキストファイルが特定できるか実験を行い，3 文字程度でも複数の語があれば特定しやすくなることを確認した．

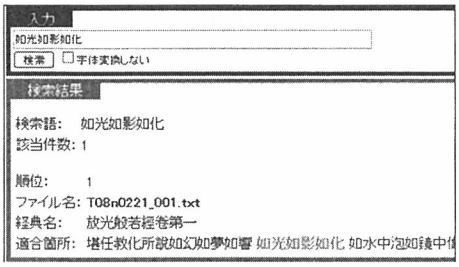
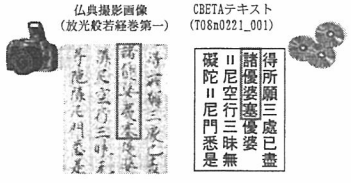


図 1　漢訳仏典全文検索システムの画面例

2.3 検索の失敗例

前節で紹介した仏典全文検索システムを使って，検索語を入れて検索しても，期待する文章が取得できないことがある．全文検索システムは N-gram に基づく漏れのない検索を実現しているので，検索失敗の原因として，画像から取り出した検索語と全文検索システムの中に保存されたテキストファイルの内容の違いが考えられる．内容の違いは，字体の違いと，誤字・脱字，意図的な変更に分類できる．字体の違いとその対策は，[4]で検討している．

脱字の例を一つ挙げる．画像側は“諸優婆夷塞優婆”であるのに対して，テキスト側は“諸優婆塞優婆”である（図 2）．テキスト側は画像側より“夷”という文字が抜けているため，そのままで検索しても，テキストファイルを取得できない．



ゲタ記号は元の CBETA テキスト（XML 形式）における外字を表す

図 2　仏典画像とテキストの異同の例

本稿で述べる「仏典テキスト校訂支援システム」は，この誤字・脱字，および意図的な変更を発見し，その情報を登録して，画像から全文検索可能にすることを目指している．

3. Subversion を用いた仏典テキスト校訂支援システム

3.1 校訂とは

全文検索システムの検索精度を向上させるため，経典画像の内容と合うように，テキストデータを修正する作業が必要である．本稿ではその修正作業を「校訂」と呼ぶ．

校訂作業の流れを図 3 に示す．画像とテキストファイルを見比べ，両者に異同が見つかれば，画像を参照しながら，テキストファイルを修正する．本研究では，このような校訂作業を支援するシステムを開発した．

なお，本研究における校訂では，全文検索システムの検索精度を上げるため，画像とあったテキストファイルを修正することを目指している．そのため，常に画像の記述内容に従って校訂を行うことにする．たとえ画像側の記述が意味的に不適切であるとしても，画像に合わせたテキストファイルの作成を支援することにした．

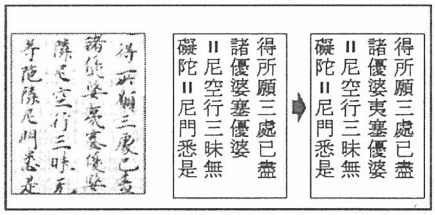


図 3　テキスト校訂の例

「校訂」とよく似た概念に「翻刻」がある．校訂とは，書物の本文を異本と照合したり，語学的に検討したりして，よりよい形に訂正すること（大辞泉による）である．翻刻とは，写本・版本などを，原本どおりに活字に組むなどして新たに出版すること（同）であり，国内では翻刻支援システム開発プロジェクト HCR[5]が有名である．古文書に合ったテキストデータを一から作る翻刻作業とは違って，本研究で述べている校訂とは，仏典撮影画像とそれに対応しているテキストデータが既にある状態で，両者を比べながら，テキストデータの異同がある箇所を修正する作業である．したがって，校訂は翻刻と比べてかける手間は少ないと言える．

文学の分野で「異同」とは，写本・口述筆記する際の写し間違い，作者や他の人間による改訂など様々な理由で派生した，それぞれ微妙な違いを持った同一作品のバリエーションのことを指す．ここで「異同」に対応する，情報工学における概念として，「差分」を取り上げたい．

2 つのテキストファイルが与えられたとき，行ごとの中身の違いだけでなく，一方のテキストファイルにのみ含まれている行も求めるアルゴリズムが存在し，それを用いて差分を出力する Unix 由来のコマンドは diff と呼ばれる．本研究でもこの意味の差分を用いて，テキストファイルの修正前後の情報を保持し，表示することとする．

3.2 Subversion

テキスト校訂作業において，どこをどのように変更するのかを管理することが重要となる．

そこで，ソフトウェア開発の分野でよく用いられる「バージョン管理システム」を活用することにした．その中でも，クライアント／サーバ型の Subversion[6]を採用する．

Subversion を導入するメリットは以下の通りである．すなわち，任意のタイミングでバックアップでき，安心して使える．そして，校訂内容だけでなく，変更記録（コミットログ）や変更者，変更日時などの校訂過程を記録に残せる．さらに，複数の作業者の共同作業をサポートできる．図 4 に， Subversion を用いた仏典テキスト校訂システムの利用形態を示す．ここで，Subversion に関する用語を説明する．サーバで管理しているファイルを「リポジトリ」と呼ばれ，「作業コピー」は，リポジトリに結び付けられた，クライアント側で変更可能な領域を指す．「アップデート」はリポジトリから作業コピーへ最新のファイルを送る作業であり，「コミット」は反対に，作業コピーで行った修正をリポジトリへ送って反映させる作業をいう．リポジトリの変更は「リビジョン」という通し番号で管理され，コミットにより，リビジョンは 1 増える．Subversion はしばしば SVN と略され，小文字の svn は，各種操作のコマンド名である．Subversion を活用するには，他に「チェックアウト」という作業や，「競合」の概念も不可欠であるが，本稿では省略する．

各ユーザは Web ブラウザを介し，Scroll Viewer[7]の機能強化版により画像とテキストを同時に参照し，テキストの修正を行う．ただし，上述のアップデートおよびコミットは，SVN サーバと Web サーバとの間で行う．また，リポジトリから取り出した各ユーザの作業コピーは Web サーバに保存し，ユーザは持たないものとする．これにより，ユーザは作業コピーを管理する手間を省くことができる．データベースを用いた Web サービスにおいて，ユーザは DBMS や SQL を認識することなく，情報を操作したり取得したりできるのと同様に，本システムでも，ユーザは Subversion の各コマンドを知らなくても，テキスト内容を変更したり，比較したり，古い内容を取得したりできる．

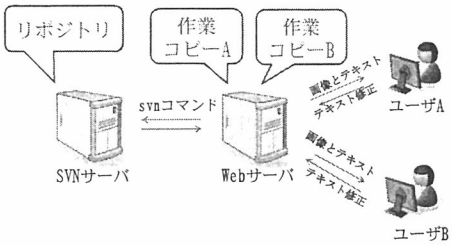


図 4　仏典テキスト校訂支援システムの利用形態

3.3 サーバのファイル管理

サーバの持つファイル構成について検討する．すでに述べたとおり，サーバ側で各ユーザの作業コピーを持ち，別々に仏典テキストファイルを校訂できることとする．また，SVN サーバと Web サーバは同一の計算機とする．

図 5 にファイル構成を示す．複数人が別々に仏典テキストファイルを校訂できるように，user/の下に複数ユーザのスペースを設けた．各ユーザが編集するテキストファイルと別に，校訂可能な仏典テキストファイルを効率よく求められるよう，リポジトリ全体の作業コピーを持つ隠しユーザ All を設けた．図 5 において，作業コピーのディレクトリは All と T08n0221_001 の二つだけである．

リポジトリには file://から始まる URL（リポジトリ URL）でアクセスする．例えば，作業コピー上のファイル /opt/ButtenRevise/user/dingmin/T08n0221_001/butten.txt は， file:///opt/ButtenRevise/repository/T08n0221_001/butten.txt という URL に対応する．

/opt/ButtenRevise 以下のすべてのディレクトリおよびファイルの所有者は，WWW サーバ Apache の実行ユーザ(www-data)としている．

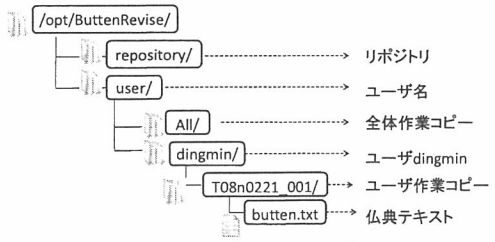


図 5　サーバのファイル構成

3.4 画面例

3.4.1 校訂テキストファイルの選択

以上の検討により，仏典テキスト校訂支援システムを実装した．その画面例について述べる．

ユーザはブラウザを起動し，指定の URL にアクセスする．認証画面（省略）であらかじめ与えられたユーザ名とパスワードを入力して

[ログイン]ボタンを押すと，仏典テキスト選択画面（図 6）に移動する．

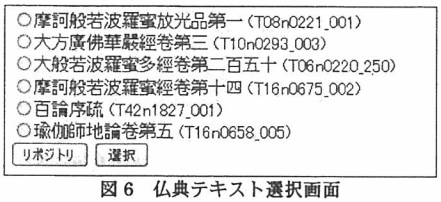


図 6　仏典テキスト選択画面

図 6 の画面では，ユーザが作業コピーとして持つすべての仏典テキストファイルの名前が表示されている．ここでユーザが T08n0221_001 の仏典を選び，[選択]ボタンを押すと，仏典詳細画（図 7）に移動する．

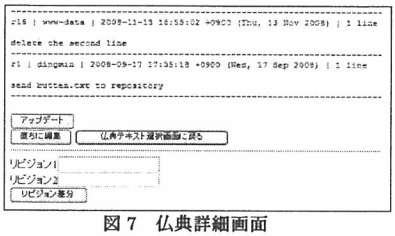


図 7　仏典詳細画面

図 7 の上部では，リポジトリが管理している T08n0221_001 について，すべてリビジョンとコミットログが表示されている．ユーザがリビジョン情報の直下の[アップデート]を押すと，アップデートを行ってから，[直ちに編集]ボタンを押すと，アップデートを行わずに，校訂の画面に移動する．図 7 の下部は，任意の二つのリビジョンの差分を確認するためのフォームであり，3.4.3 節で述べる．

3.4.2 校訂画面

校訂の画面は，閲覧画面（図 8）と修正画面（図 9）の二つがあり，相互に移動できる．図 8 は，Scroll Viewer を拡張したものである．画面の左右には，T08n0221_001 の仏典画像とそれに対応するテキストが表示されている．さらに，経典画像に対してマウスのドラッグ操作を行うと，画像が動き，それに同期してテキストも対応する箇所を表示する．この機能により画像を閲覧しながら，テキスト内容との比較を行う．図 8 で，画像側とテキスト側にそれぞれ四角形で囲っているように，画像側の“諸優婆夷塞”に対応しているテキスト側の内容は“諸優婆塞”となっており，テキスト側で文字が足りないことが分かる．このときユーザがテキスト下の[訂正する]というボタンを押すと，修正画面（図 9）に移る．図 9 では，テキストフォームによりテキストを任意に書き換えることができる．実際，ユーザがテキスト側の“婆”と“塞”の間に“夷”を追加し，テキスト下の[更新する]ボタンを押せば，再び閲覧画面（図 10）に戻る．ここでは，画像側とテキスト側ともに“諸優婆夷塞”の内容になっている．

この時点では，修正はローカルの作業コピーに対して行われており，十分に修正を行ってから，その内容をリポジトリに送信するコミット作業を行う．それには，ユーザが仏典画像上の[コミット]ボタン（図 8，図 11）を押せばよい．競合が起こらなければ，図 11 のように，テキスト側の直下に「コミットします」というメッセージが表示される．

図 8　仏典閲覧画面

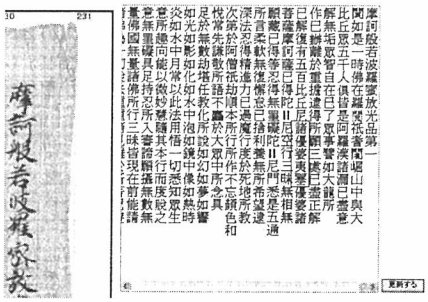


図 9　仏典テキスト修正画面

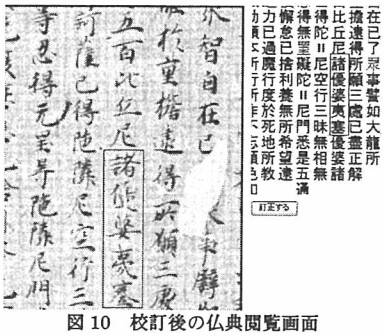


図 10　校訂後の仏典閲覧画面

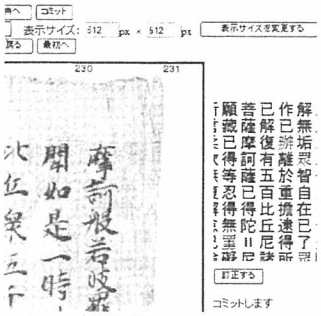


図 11　仏典閲覧画面でコミット

3.4.3 リビジョン間の差分

リビジョン間の差分を確認する手順について述べる．3.4.1 節でも少し触れたように，仏典詳細画面はリポジトリが管理している仏典テキストファイルの全リビジョンが表示される．ここで図 12 のように，リビジョン記入欄に，ユーザが 34 と 35 を入力して，[リビジョン差分]ボタンを押すと，それらのリビジョンの相違点を見ることのできる差分画面（図 13）に移動する．この画面から，3.4.2 節で例示した修正に成功したのが確認できる．

図 12　仏典詳細画面

```
--- butten.txt  (revision 34)
+++ butten.txt  (revision 35)
@@ -3,7 +3,7 @@
 比丘衆五千人俱皆是阿羅漢諸漏已盡意
 解無垢衆智自在已了衆事譬如大龍所
 作已辦離於重擔逮得所願三處已盡正解
-已解復有五百比丘尼諸優婆塞優婆夷諸
-已解復有五百比丘尼諸優婆塞優婆夷諸
 菩薩摩訶薩已得陀＝尼空行三昧無相無
 願藏已得等忍得無量辯陀＝尼門悉是五通
 所言柔軟無復僑慠已捨利養無所希望遠
@@ -444,4 +444,4 @@
 種力四無所畏及佛十八法乃至薩云若法
 亦不見應與不應是故舍利弗當知菩薩與
 般若波羅蜜相應
```

図 13　差分画面

4. 全文検索システムの改良

大正新脩大蔵経テキストデータの動向について，この 1 年で特筆すべき事項として，SAT がオンライン検索の提供を始めたことが挙げられる（http://21dzk.l.u-tokyo.ac.jp/SAT/database.html）．行番号有無の選択や，注釈をボタンで表示可能など，より柔軟に大正新脩大蔵経を「読む」機能が提供されている．他サイトからのリンク方法を定めている点も興味深い．

他の仏典テキストの現状は[8]にまとめられており，本研究で対象とする奈良平安古写経の電子化についても課題として指摘されている．

本研究では，全文検索システムおよび大正新脩大蔵経テキストデータは[4]で構築したものに基づき，機能拡張を行った．そこで解決すべき点は，複数リビジョンに配慮した検索である．しかし，最新リビジョンのみを登録するのでは，途中の修正分が検索の対象とならない．また，単純に各リビジョンのテキストファイルを別個に登録するのでは，検索結果の該当件数が「経巻数」ではなく，「ファイル数」となってしまい，利便性を損なう．

そこで，各リビジョンのテキストデータを登録し，検索において，同一経巻文書に複数出現する場合には，その中の最新のリビジョンの文書を提示することを試みた．仏典テキストはデータベースで管理し，その管理システム(DBMS)には PostgreSQL を，全文検索エンジンには Senna および Ludia を採用した．仏典テキストの情報を格納するテーブルは，本文（属性名 body）だけでなく，CBETA に準じた文書名（同 filename），リビジョン（同 revision）に関する属性を持つ．さらに，属性 body に関しては N-gram で情報を格納し全文検索を行うためのインデックスを，また filename および revision については通常のインデックスを生成するよう指定している．[4]で生成した 8,989 件の各経典テキストファイルをリビジョン 0 として登録し，いくつか SQL 文を与えて検索を試みたところ，これまでの検索と同じヒット数および文書の取得ができた．また Senna, Tritonn, MySQL を用いた従来版では，1 文字の検索に時間を要するものがあった（ただし，DBMS の違いというよりも，インデックスや SQL 文に問題があった可能性が高い）が，本システムでは 0 件であっても数千件であっても，件数を瞬時に求めることができた．

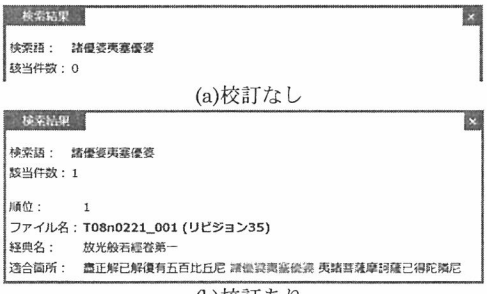
リビジョンを考慮した文書取得については，放光般若経巻第一(T08n0221_001)のリビジョン 0 のテキストデータに対して，末尾から 1 行ずつ順に削除して新たなリビジョンとして（ユーザによる校訂と区別するために，revision は 1,000 から）登録し，以下の SQL 文で，最大のリビジョン番号を持つ経典を検索した．

```
SELECT c1.filename, c1.revision
FROM canons c1
WHERE c1.body @@ '*D+ 所言柔 深法忍
次第於' AND NOT EXISTS(SELECT c2.id
FROM canons c2
WHERE c2.body @@ '*D+ 所言柔 深法忍
次第於' AND c1.filename = c2.filename
AND c1.revision < c2.revision);
```

瞬時に求めることができ，内容を見て正しい

ことを確認した．検索語として任意の Senna 検索式が利用可能であるが，「菩薩」のような頻出語では，数秒かかった．この問題については，利用者にストレスを感じさせない閾値を設け，ヒット数がそれ以下ならば同一で複数のリビジョンの文書は最大のものを求め，そうでなければ同一で複数のリビジョンの文書も別々に求めるといった運用方法が考えられる．

以前の検索システムでは，経典画像から「諸優婆夷塞優婆」を取り出し，検索を行っても，期待するテキストファイルが取得できなかった（図 14(a)）．しかし，上述の全文検索システムおよびテキストファイル登録機構を構築し，画像に沿ってテキスト内容を修正してコミットしてから，同じ文字列で検索を行うと，成功することを確認した（図 14(b)）．



(a)校訂なし

(b)校訂あり

図 14　校訂を考慮した全文検索

5. おわりに

本稿では，仏典撮影画像に合わせたテキストファイルの作成を目的として開発した，バージョン管理システム Subversion を用いた仏典テキスト校訂支援システムについて述べた．その特徴は，(1)「異同」と，情報技術における「差分」を結び付けたこと，(2) DBMS を用いた Web サービスと同様に，Subversion のコマンドなどをサーバに隠蔽したこと，である．なお，差分を DBMS ではなく Subversion で管理したのは，任意の時点（リビジョン）のテキストファイルを取得できるだけでなく，任意の 2 つのリビジョンの同一テキストファイルの差分を容易に取得できるためである．

画像以外（対校本テキストなど）と照合して校訂するインタフェースや，Subversion におけるブランチやマージなどの機能をシステムに取り入れ，より多くの仏教学などの専門家に利用してもらい，修正を重ねていくことで，[9]で提唱されている「21 世紀大正蔵」を補助するシステムになり得ると考えている．

謝辞　本研究を進めるに当たり，様々なご教示を賜った，国際仏教学大学院大学 落合俊典教授に心より感謝いたします．

参考文献

[1] 落合俊典: 金剛寺一切経の総合的研究と金剛寺聖教の基礎的研究 平成 16～18 年度科学研究費補助金基盤研究(A)研究成果報告書, 課題番号 15202002, 第 1 分冊 (2007).
[2] 落合俊典: 金剛寺一切経の基礎的研究と新出仏典の研究, 平成 12 年度～15 年度科学研究費補助金基盤研究(A)・(1)研究成果報告書, 課題番号 12301001, (2004).
[3] 張蓉, 仁野洋平, 田中猛彦, 中川優, 青木進, 宇都宮啓吾, 落合俊典: 仏典データベースのための画像処理について, 情報処理学会研究報告, 2006-CH-69, pp.25-32 (2006).
[4] 村川猛彦, 丁敏, 中川優: 仏典全文検索システムの構築と評価, じんもんこん 2007, 情報処理学会シンポジウムシリーズ Vol.2007, No.15, pp.221-228 (2007).
[5] 翻刻支援システム開発プロジェクト HCR, http://www.nichibun.ac.jp/~shoji/hcr/
[6] Subversion, http://subversion.tigris.org/
[7] 松浦康夫, 経典画像閲覧システムにおける画像と文書の対照表示機能, 2006 年度和歌山大学システム工学部卒業論文 (2007).
[8] 石井公成: 大蔵経データベース, 漢字文献情報処理研究, Vo.9, pp.149-153 (2008).
[9] 落合俊典: 漢訳仏典研究の新たな視座―日本古写経のデータベースと SAT&CBETA の利用―, 国際仏教学大学院大学学術フロンティア公開シンポジウム講演資料集, pp.77-87 (2007).