PDFファイルを用いた歴史的資料のデジタル化とデータベースの試作

An Approach on Interactive Database using Annotation of PDF Documents

竹内 さおり　　白川 哲郎
TAKEUCHI Saori　　SHIRAKAWA Tetsuroh

大阪樟蔭女子大学　学芸学部，東大阪市菱屋西 4-2-26
Osaka Shoin Women's University, 4-2-26, Hishiyanishi, Higashiosaka, Osaka
takeuchi.saori@osaka-shoin.ac.jp

あらまし：本稿では，樟蔭学園デジタルアーカイブズ構築の一部として，昭和戦前期の学園広報誌『樟蔭學報』のデジタル化とデータベース試作の取り組みについて報告する．『樟蔭學報』は，1936年（昭和11年）8月に創刊され，創刊号から第三巻第三号までの合計19冊が現存する．本研究では，資料をデジタル化して保存・整理することとこれらを授業で活用することの二つを目的とする．授業で活用するための工夫として，データベース本来の検索や検索結果を閲覧する機能に加え，内容の充実に利用者が参加できるような仕組みを考えた．具体的には，PDFファイルの注釈を用いて，既存のキーワードに利用者の保持する情報やコメントを付加することを可能にし，インタラクティブな仕様を実現した．

Summary: In this paper, I will report on digitalization and the making of database for the magazine 'Shoin Gakuhou', for the first term of the war of the Showa era. This project is a part of constructing the educational institution digital archives of Shoin Gakuen. 'Shoin Gakuhou' started in August, 1936, and presently, there are nineteen books consisting of volume 1 to volume 3, the third issue. There are two aims for digitalizing these magazines to preserve and arrange them, and to use them in the classroom. In addition to the user-friendly interface of accessing information, the program of the database is designed for student users to add new information and comments easily to the contents. They are expected to type in their own information under existing keywords of the data. This interactive specification was achieved by using the annotation function of PDF documents.

キーワード：デジタルアーカイブズ，PDFファイル，注釈，データベース，双方向性
Keywords: digital archives, PDF files, annotation, data base, interactive

1.はじめに

1917年（大正7）に樟蔭高等女學校として開学した樟蔭学園は，2007年に創立90周年を迎える．学園草創期の資料は，近現代の女性史及び学校教育史を考察するうえで貴重な資料であるとともに，大阪樟蔭女子大学大学史を理解するための重要な教材である．

平成15年度より継続的に進めてきた樟蔭学園草創期資料のデジタル化とデータベース化に関する研究は，作業手順をほぼ確立することができ，進捗状況は順調である．着手から3年目にあたる平成17年度は，保存と整理の作業そのものよりもデジタル化した資料やデータベースを教材として活用することにポイントをおき，使い易さや楽しさをとり入れる工夫をした．1936年（昭和11）創刊の学園広報誌である『樟蔭學報』を対象とし，デ

ジタルコンテンツの利点を生かし，知識の共有も容易に行える教材の作成に取組んだ．

本稿では，PDF(Portable Document Format)[1] ファイルの注釈機能を用いたインタラクティブなデータベースの試作について報告する．以下，2節で研究の背景と目的を述べ，3節で提案手法を詳細に説明する，4節で『樟蔭学報』の内容にふれ，5節で基本作業手順と具体例を示す．6節では評価実験について述べ，7節で関連研究を挙げる．最後に，考察と今後の予定を述べる．

図 1：『樟蔭學報』創刊号の表紙

2.研究の背景と目的

本研究は，樟蔭学園の歴史的資料を授業で活用することで，学生の教育効果を期待するとともに母校への愛着を増加させたいという筆者等の考えからスタートした．平成 16 年度には，データベースソフト(Microsoft Access)を用いた『設立二関スル書類』のカード型データベースを試作し，紙資料のデジタル化とデータベース化を行った[2]．しかしながら，テスト運用の段階で以下のような指摘を受けた．

1. 1 枚のカードで資料 1 ページ分の画像と各項目のテキストデータを見ることができるが，資料全体のイメージは把握しにくい．
2. 文字や画像が小さすぎて見えにくいので，教材としては相応しくない．
3. 画像をそのまま複製して転用される可能性があるので，データでの配布や Web で公開することは好ましくない．

これらの指摘を改善するために，データベースソフトに依存することや画像データとテキストデータを並べて表示する型に固執せず，コンピュータの画面上でパラパラとページを繰るように閲覧できる仕組みの実現と検索の容易さ，データでの配布や Web で公開する際に不都合が生じないように配慮することを目的に，新たな手法を検討した．

3.提案手法

本研究では，PDF ファイルの特徴に着目し，データベース構築への利用を検討した．

PDF ファイルは，プラットホームや使用環境に影響を受けず，利便性の高い文書フォーマットであり，文書の配布に PDF ファイルを利用する頻度も高くなっているので，利用価値が高く有用なファイル形式とあると考えられる．また，PDF ファイルは，文書ファイルからだけでなく，画像ファイルからの変換も可能であり，変換後のファイルに遜色がないばかりかファイルサイズが小さくなるためネットワーク上で扱い易い．検索に関しては，既存の全文検索機能が利用できるが，文書ファイルから変換した場合に限られるため，OCR 装置やデジタルカメラで入力したデータは対象外とされていた．

今回のように，画像ファイルから生成した PDF ファイルの文字列を検索する方法として，注釈機能を利用した手法を提案する．つまり，キーワードを注釈として追加し，注釈を手がかりに PDF ファイルの中身を検索しようという発想である．

注釈を利用するもう一つの理由は，データベースの作成者だけでなく，利用者も注釈を追加することに参加できるからである．

具体的には，以下のような方法である．

- PDF ファイルの作成者がつけたキーワード

を用いて検索した利用者が新たなキーワードを注釈として追加する．

- 閲覧した利用者が重要と思われる情報や関連した資料についての注釈を追加する．

利用者が注釈を追加することでキーワードを増加することができる．さらに，その人が何を重要と考えているかも明らかになる．

つまり，利用者が積極的に参加できる環境を整えることができれば，知識を共有するという観点からも効果が期待できる．例えば，図書館で借りた本を教材として利用していたときに，重要だと思ったことや他に調べた資料のことを直に書き込むことはできない．他の利用者が何を理解して何を重要と考えたかという見えない情報や参照した関連資料の記録は明らかにならない．

しかしながら，このような情報が役立つ機会は多いように思われる．紙の上では実現できないならば，PDF ファイル上で実現したいと考えた．

提案手法では，Amazon のお薦め情報の提示[3]に類似した状況を機械的なアルゴリズムからではなく，人と人のつながりから実現する．

教員と学生及び学生同士のコミュニケーションツールの役目を果たす可能性も高く，教育効果も期待できる．教材を介し知識を共有できる場を提供するとともに，注釈の蓄積は原資料に付加価値を付けることもでき，有用性も増加する．

一方，画像の複製や転用に関する件は，PDF ファイルを作成する際にセキュリティ設定を変更することにより防止が可能であることを確認した．

利用者の編集や印刷の操作に制限を加えることが可能になるので，画像データであっても保護することができる．データでの配布や Web で公開する場合に不安とされた点が解消でき，ネットワークを介した遠隔授業等での利用も予定している．

従来のデータベースのように，蓄積したデータを利用するだけでなく，動的な仕様[4]を実現することも可能かもしれない．

4.『樟蔭学報』について

『樟蔭學報』は 1936 年（昭和 11）に創刊された樟蔭学園広報誌であり，創刊号から第三巻第三号までの合計 19 冊が現存する．創刊号掲載の「樟蔭學報発刊の辭」には，在学生二千名，樟蔭高等女學校本科・専攻科・高等科の卒業生同窓会員（緑蔭会員）三千数百名，樟蔭女子専門學校卒業生同窓会員（緑翠会員）一千数百名，それぞれの保護者を含めた合計一萬数千名の関係者に向けて，親交を温めるために発行されたと記載されている．内容については，樟蔭高等女學校及び樟蔭女子専門學校の年中行事の紹介，経過報告，緑蔭会及び緑翠会の会記や会員の消息，生徒の詩歌や文章の発表，緑蔭会員及び緑翠会員の作品発表，保護者の声と学校側の反響，論説，文苑等となっている．行事や校内で撮影された写真も多く掲載されていて，服装や行事の様子などもよくわかる．当時の様子がうかがえる広告などもたくさん含まれ，非常に興味深い資料であると思われる．なお，図 1 に示した創刊号の表紙は，カラー印刷である．

5.基本作業の手順

図 2 に基本作業の流れを示す．基本作業は，二つの部分から成る．第 1 段階では，『樟蔭學報』の各ページをデジタルカメラで撮影して画像を保存し，学園資料調査カード（紙媒体）に必要な項目を記入する．画像データとテキストデータを 1 つずつ作成していく地道な作業である．第 2 段階では，画像データを修正し，PDF ファイルに変換してから，画像の撮影時に抽出したキーワードを注釈として付加する．今回の作業では，各号ごとに PDF ファイルを作成した．第 1 段階は白川ゼミ卒業生の松田憲子さんに担当を依頼し，第 2 段階は竹内が担当した．本稿執筆の段階で，創刊号からの第一巻五号までの 5 冊分が完成し，現在も作業を継続中である．以下に，詳細な説明と具体例を示す．

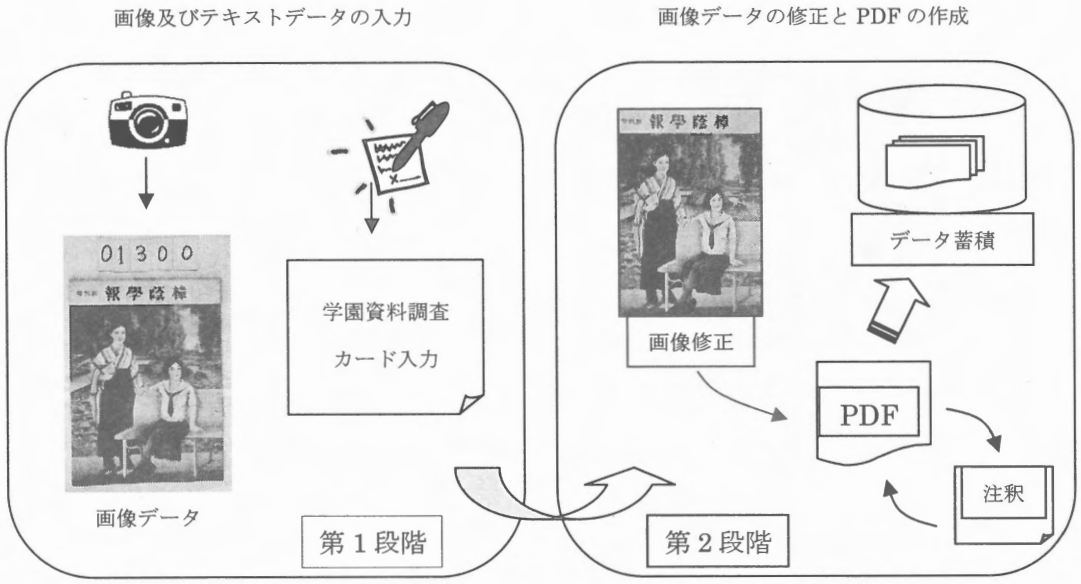


図 2：基本作業の流れ

5.1 画像の撮影と学園資料調査カードの記入

画像は，1 ページずつ順番にデジタルカメラで撮影した．デジタルカメラの誤操作によるデータの消失を避けるために CD 内蔵のデジタルカメラ[5]を使用した．学園資料調査カードの様式は，前回と全く同じ A4 サイズの用紙に「カード No.，調査日，調査者，名称，資料名，形状，材質，サイズ，年記，内容，キーワード，写真，特記事項，学園整理台帳における記載の有無，記載者」の 15 項目で作成したものを使用した．学園資料カードの記入（手書き）と Microsoft Excel のワークシート入力の両方を行うことに，冗長的な感じは否めないが，作業方法の変更は行わない．第 1 段階で特筆することは，撮影後にハードディスクへ保存する方法でデータの二重化につとめていたにも関わらず，ハードディスクのクラッシュによりデータの一部を消失してしまった点である．解決策としては，毎日決まった時間に別のハードディスクにバックアップを作成するタイプ[6]に更新した．データの消失は，保存データが増加すればするほど，大きな痛手となる．今後も，データの管理やバックアップの方法については，その都度，信頼できるハードウェアを採用していくこと等で対応したい．

5.2 画像の修正と PDF ファイルの作成

図 3 に PDF ファイルの例を示す．デジタルカメラで撮影した jpeg 形式のファイルは，コントラストと色調の微調整を行い，画質の良い状態で保存し直した後，1 冊分ずつまとめて PDF ファイルに変換する．まとめて変換することで，1 枚ずつバラバラに撮影した画像をひとまとまりとして扱うことができるので，1 冊の本のようにパラパラと閲覧することが可能となる．画像の修正には Adobe Photo Shop を PDF ファイルの作成には Adobe Acrobat を使用した．PDF ファイルに変換した後に，キーワードとして抽出した語を注釈（ノート）の追加を行う．

5.3 注釈を用いた検索と拡張性

図 4 に注釈を用いた検索の例を示す．「旅行」をキーワードに創刊号を検索したところ，3 件の結果が得られた．該当する「東京地方修学旅行」のページと注釈が表示され，注釈のマッチした部分が太字になる．ここでの検索は，注釈として付加された語だけを対象にしているので，検索文字列と全く同じもしくは一部が，注釈と同じ文字列の検索にすぎない．ここでは，関心のある箇所を閲覧

図3：PDF版『樟蔭學報』の例

し始める入口の意味合いで検索結果を捉えたい．前述したように検索で得られた結果が最終目標ではなく，利用者が注釈にキーワードを追加することやコメントを書き込むことによって，知識を共有し再利用するための手がかりである．

例えば，演習形式の少人数のクラスで使用する場合，ネットワーク上のPDF版『樟蔭學報』に同時に複数の学生が各自のコンピュータからアクセスして，閲覧しながら授業に参加する．授業時間内にキーワードや個々のコメントを注釈として追加することができなかったとしても，授業時間以外にも同じように利用できるので，時間的な制約も受けない．教授者はコメントに対するアドバイスを与え，議論の活発化をはかる．

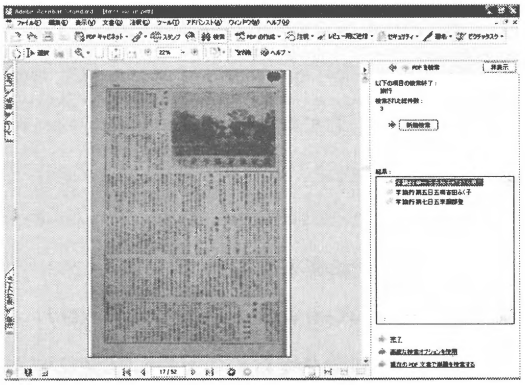


図4：注釈を用いた検索の例

5.4 作業の進捗状況

創刊号から第三巻三号までの合計19冊の撮影と学園資料カードの記入が完了し，テキストデータの入力と最近になって見つかった緑蔭会員[7]及び緑翠会員[8]向け冊子の特集記事を撮影中である．画像の修正が終了したものから順にPDFファイルへの変換作業を済ませているので，テキストデータが揃えば注釈付けは完了できる．

6.評価実験

本研究における成果を平成18年度に白川が担当する授業で活用する予定であり，現段階で評価実験を行った．以下では，評価実験の方法，被験者の反応とアンケートの集計結果について述べる．

6.1 評価実験の方法

被験者には，大阪樟蔭女子大学学芸学部日本文化史学科1回生8名と2回生9名に協力を得た．「基礎ゼミ」の授業時間に，研究の主旨や目的を説明し，実験に協力してもらった．まず，冊子体の『樟蔭學報』とPDF版『樟蔭學報』の両方を見せる．前者は，手にとって自由に触れられるようにし，後者はコンピュータの画面をテレビに映して，キーワードによる検索の様子と注釈を追加する方法について，操作を交えて説明した．実際に，情報やコメントを追加して，注釈を複数の利用者で共有して活用できる仕組みについて説明した．30分間程度，コンピュータの操作を試した後に，簡単なアンケートに回答してもらった．

6.2 被験者の反応

同じように実験を進めたにも関わらず，1回生のクラスと2回生のクラスでの反応は全く異なった．1回生のクラスでは，関心を示す者が多く，非常に活発な議論がなされた．一方，2回生のクラスでは冊子体の『樟蔭學報』にもコンピュータの画面や操作のいずれにも，関心を示す者が少なく，議論も殆どされなかった．以下に，1回生のクラスで議論された内容を列挙する．

- 校歌の楽譜のページで，クリックしたら校歌が聞けるようにしたらいいと思う．

- ◆ 校歌を歌う機会も少ないし，歌えないから．
- ◆ このページの写真の建物が現在はどうなっているかが見られると楽しい．
- ◆ これって，正門のとこの建物？
- ◆ 伊賀駒吉郎って誰れ？
- ◆ 校長先生って書いてある．
- ◆ この本が発行された頃はどんな時代だったんですか？
- ◆ 二二六事件のあった年です．（教授者）
- ◆ 時代背景などがわかるように，年表と一緒に表示できると面白い．
- ◆ 広告の文章が面白い．
- ◆ 蚊が取れるから，カトール？

6.3 アンケートの結果

以下にアンケートの集計結果を示す．

アンケートの質問は，予め回答群を用意した3問と自由回答1問の合計4問である．

1. 今日のようにデジタル教材を使う授業について，どう思いますか？【複数回答可】

回答群	人数
面白い	15
面白くない	0
楽しい	7
楽しくない	0
積極的に参加したい	5
あまり参加したくない	0
難しそう	2
操作が面倒	2
現物(本)のほうが良い	3
現物(本)より扱いやすい	8

2. 樟蔭学園の歴史に関心がありますか？【選択】

回答群	人数
ある	1
少しある	13
どちらかというと無い	3
全く無い	0

3. デジタル教材に書き込みをして，知識を共有することについてどう思いますか？【選択】

回答群	人数
関心をもった	9
少し関心をもった	8
全く関心が無い	0

4. 自由な感想を何でも【自由記入】

- もっとデータの数が増えるのがたのしみ*です．
- パソコンを使って自分の知りたいことを一瞬で調べられるのは大変ベンリ*だと思いました．
- 昔のことが分かってよかったと思う．
- デジタル教材を積極的に使ってほしいと思いました．
- 私は本を読むことが好きなので現本*を読むほうが好きですが，最近はマンガもデジタルで読む時代になっているので，どんどん作るべきだと思います．
- 検さく*がしやすそうです．
- 少し画面が見にくかったので，見やすくなればいいと思いました．
- パソコンでデジタル教材として扱うのはすごいと思った．
- どんどん新しい教材ができるのは賛成です．
- 樟蔭の歴史を知ることができるしおもしろそうだと思います．
- 現物より，多くの人が見れるのは良いと思いますが，正直*見やすさは本の方が上*だと思います．
- どういう授業になるのか一度体験してみたい．
- デジタルなら本物のように壊れやすくないので，積極的に活用できそうだと思います．
- キーワードを入れたら調べたいもの*がでてくるというのは便利

* 自由回答の箇所については，被験者が記入した表記のまま転載した．

- 他の資料でも，デジタル教材にしたらおもしろそうだと思いました．
- 校歌の所*に曲がながれるようにしてほしいです．

7.関連研究

國學院大學日本文化研究所で進められている中村等[9]の研究は，デジタルアーカイブシステムパッケージ「でんとうなび®」[10]を利用したWeb版データベースの公開である．國學院大學が所蔵する考古学・民俗学・国文学・文化財学・神道学等の分野の重要な業績と膨大な研究資料の写真資料を中心に，冊子体目録・CD-ROM版データベース・Webデータベースの構築を実現した．ユーザー用メモ欄を設けるなどの工夫が参考になった．

また，大橋[11]の南山大学図書館における収蔵の貴重書『ローマ法大全』をデジタルアーカイブ化に関する報告では，映像を交えた資料紹介のWebページ作成について詳細に述べられている．なかでも，エンドユーザーの立場から見て利用しやすい，かつ独創的な発信内容を満載した利用型のアーカイブ構築したい旨の将来構想には，共感するところが大きい．

多様な資料構造に対応したデジタルアーカイブシステムについて，依田等[12]はメータデータツリーの生成という手法で，検索の柔軟性と拡張性を実現した．利用者の要求に即した情報を網羅的に収集するためには，章や節といった構成要素の単位まで検索できなければならない．検索者の意図とメタデータ間の関係を利用して，検索方法を使い分ける仕組みに注目した．

8.考察と今後の予定

本稿では，昭和戦前期の学園広報誌『樟蔭學報』をデジタル化し，PDFファイルの注釈機能を利用したインタラクティブなデータベースとして作成し，教材として活用することについて述べた．

先行研究[2]におけるカード型データベースのテスト運用で，指摘された問題点を改善し，従来型の画像データとテキストデータを並べて表示する型式に固執せず，コンピュータの画面上でパラパラとページを繰るように閲覧できる仕組みと利用者が内容の充実に参加できるインタラクティブな仕様を実現した．

さらには，データで配布する場合やWebで公開を想定して，データのセキュリティ面の強化にも努めた．

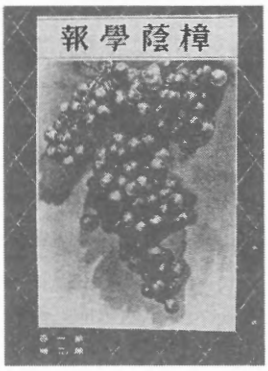
今後は，残り14冊分の注釈を早急に作成し，利用者側の環境を整備した後，実用に備えたいと考えている．

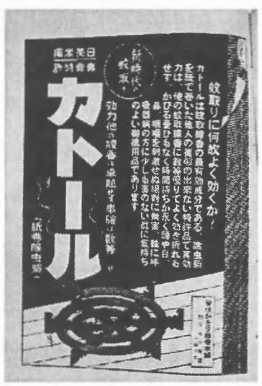
今後，検討しなければならない事項として，以下のような点を挙げておく．

1. ネットワーク上においた同一ファイルに複数の利用者が同時に注釈を付加したときに，更新が正常に行われるか検証する必要がある．
2. データの増加とともにバックアップの作成は必須であり，とくに慎重に取り組みたい．
3. 1冊ずつ個別にPDFファイルに変換したが，1冊単位としたことで検索範囲を狭めてしまい，検索結果を得にくくなっているので，19冊全てを1つのPDFファイルとして作成し直すことも検討する．
4. コンピュータの操作に不慣れな利用者を支援するためのマニュアルづくりにも対応したいと考えている．
5. PDFファイルの検索には，インデックスをつける方法やNamazu[13]による全文検索の方法などが紹介されている情報[14]も得られたので，今後の検討材料としたい．

付録

第一巻第二号から第五号の表紙と創刊号の広告





謝辞

本研究の遂行にあたり，多数のアドバイスをいただいた大阪樟蔭女子大学学芸学部英米文学科小森道彦助教授並びに藤澤良行助教授に心より御礼申しあげたい．また，画像の撮影等データ入力の作業を担当してくれた奈良女子大学大学院人間文化研究科博士前期課程1年の松田憲子さんと評価実験の被験者として協力を得た方々に深謝する．

本研究は，大阪樟蔭女子大学平成17年度特別研究助成費によるものである．ここに記して謝意を表す．

参考文献

---

[1] http://e-words.jp/w/PDF.html

[2] 竹内さおり，白川哲郎，『設立二関スル書類』のデジタル化とデータベース試作の取り組み－樟蔭学園草創期資料のデータベース化とその活用(2)，大阪樟蔭女子大学（学芸学部）論集第42号，p213-219，2005

[3]「この本を買った人は，こんな本も買っています」という案内．

[4] 清水宏一，治田嘉明，デジタルアーカイブの実践と今後への課題，情報処理学会研究報告，2003-CH-60，p1-8

[5] CDマビカ(MVC-CD500)，SONY製

[6] Link Station(HD-160LAN)，Direct Station(HD-160U2)，BUFFALO製

[7] 樟蔭高等女學校本科・専攻科・高等科の卒業生同窓会員

[8] 樟蔭女子専門學校卒業生同窓会員

[9] 中村耕作，黒崎浩行，小川直之，杉山林継，学術調査資料の整理・公開システムの構築－写真資料を中心に－，情報処理学会研究報告 2005-CH-66，p15-22

[10] http://www.toshiba.co.jp/efort/product/digital/index_j.htm

[11] 大橋直美，貴重書『ローマ法大全』と南山大学デジタルアーカイブ，南山大学図書館紀要第8号，2003，p109-116

[12] 依田平，渡邊隆弘，大月一弘，鳩野逸生，岩杉大輔，多様な資料構造に対応したデジタルアーカイブシステム－神戸大学電子図書館アーカイブ検索システム－，情報処理学会研究報告，2003-FI-73，p45-52

[13] http://www.namazu.org/

[14] PDF HACKS，Sid Steward著，千住治郎訳，オライリー・ジャパン発行，2005

※上記URLの最終確認日は平成17年11月7日である．