第６回仏教図書館協会研修会　10月12日（金）

講演３「ＩＴ初学者の電子テキスト利用法」

花園大学文学部史学科講師　福島恒徳

はじめに

私は決して電子テキスト専門家であるとか、あるいは禅宗史、禅宗漢文の専門家ではありませんので、まずそのことをお断りしておきます。

私の専門は美術史という学問で、禅宗美術などを研究している関係で、禅宗文献に関しては美術史専攻者の中では比較的親しんでおる方かと思います。

実は、私はまだこの大学に来て２年目です。前におりました職場で雪舟や、雪舟の属する禅宗系の美術というものを専攻して展覧会をしたりしましたから、その準備にZENBASEを使っていたということがありました。ある時後藤さんや沖本先生と飲みながらお話をしていた時に、「ZENBASEを使っているのですよ、大変役に立っていて非常に助かっています」という話をしたところ、それを覚えておられまして、こういう催しがあるので、花園大学は非常に早くから文献の電子テキスト化を行い、成果を挙げているということの波及効果というか、実際どう使っているのかということをお見せするのもいいではないかということで、私に白羽の矢が立ったという訳です。

ZENBASEとの出会い

実際に私が何故ZENBASE、或いは全文テキストを使うようになったかをお話しします。師先生のお話にもありましたが、私などはDOSの時代からやっているのですが、いたる所に様々な、ほぼアングラ出版に近いような全文テキストデータベースというものが出回り始めた時期がありました。それは美術史の世界でも同様で、主として東京国立文化財研究所、最近、東京文化財研究所という独立法人になりましたが、そこで最初に所蔵雑誌の文献目録をテキストベースで発表された。そういうものを使いたくて始めたのです。その文化財研究所が、大正大蔵経と同じように大部の本を丸ごと入れたいわゆる全文テキストを、著作権の関係で１部の人間にモニターという形で利用させるようにした際には私も利用させていただきました。

そうこうする内に、私の場合はBBSとかそういうところで情報を時々入手していまして、ZENBASEが出たということを聞いた訳です。皆さんはタダでお持ち帰りになるようですが、その当時結構厳しいことを言われたことを覚えています。もう懐かしい思い出ですが、田舎の美術館の学芸員が何でそんなものがいるのだと、多分思われたと思うのですが、これは基本的に学術資料であって、資料交換という形で、実費1000円ということでお願いしますということを言われたのです。それで、私は水墨画の本とかを作ったりしていましたので、そういうものをお送りして、1000円払って、ある意味では苦労して手に入れたわけです。そして、その後本当に助かることがありました。

美術史の世界では禅宗美術というのは以前からかなり取り上げられてはいるのですが、実のところ、その領域でやっかいなものに画賛があります。水墨画とか、或いは禅宗絵画、頂相（ちんそう）などにも画賛があります。これが非常に難解なもので、ご承知の方も多

いかと思うのですが、こういうものを美術史家というものは読めないから放っておいたということなのです。例えば私などは、展覧会などで作品を拝借してきて展示をするわけですが、展覧会図録を作る時に、今申し上げたような事情があって画賛を全て起こすということをほとんどしてこなかったということがあります。これではいけない、せっかくの展覧会だし、ビシッとやってやろう、ということで、ZENBASEを使って、非常に難解な画賛を、出品作品の画賛全てを起こして載せたことがありました。「禅寺の絵師たち」という展覧会です。

例えばこういうものをご覧頂きます(吉山明兆筆「釈迦三尊三十祖像」京都市・鹿王院蔵絹本着色　掛幅装７幅　重要文化財　山口県立美術館発行『禅寺の絵師たち』展図録所載)。これはいわゆる「列祖図」です。それぞれの祖師について略歴というものが書いてある。京都鹿王院というところにあります明兆という室町初期の画僧が描いた作品なのですが、こういうものにしても実は図録類に「起こし」がなかったのです。こういうものを調べるのにZENBASEが非常に良い。今日の沖本先生のお話でもありましたが、語録が禅宗界にとってはお経のようなものでありまして、そこここに語録から語彙が引用される。そうでなくても、禅宗に限らず、漢詩などでは以前使われたレトリックを引用するのが作法というようなところがあり、これが非常に多用されるわけです。重要なフレーズというのはいろいろなところで繰り返し使われる。こちらは禅のことが全く分からないわけですから、読める部分を入力して、検索をかけて、ヒットした部分の前後を読んで、この字、このフレーズではないかという方法で解読を進めていったということです。今でもZENBASEについて私はそういう使い方をしています。

一般に利用されている全文テキストの種類

本題に入ります。コンピュータで実際に全文テキスト検索などをおやりの方には、もうほとんどそんなものは知っているというようなお話だと思います。ですが、普段、職場の都合とか、いろいろな事情があって、あまりコンピュータを使う環境でないという方もおられるでしょう。ですから、まさに初学者向けの、私自身も初学者ですが、全文テキストの意義であるとか、或いはその利用方法、その際の問題点といったようなことをお話させていただこうかと思います。

一般に流通している全文テキストには、先程の、きちんとフォーマットが決まった大蔵経プロジェクトのようなものもあれば、そうでないものもあります。例えば10年ほど前などですと、フォーマットどころか文字コードということすら知らずに、自分の持っている環境で、場合によってはワープロで作ったものを変換して使うというようなことすら行われていました。その頃からの蓄積がいまだにあるわけです。私などはそういう雑多なものを手元に置いて、それを全て使わねばならないという環境で仕事をしています。美術史の文献などというのは、悲しいかな、仏教界に比べて非常に意識が低くて、フォーマットの統一というような話にはなりにくいのです。美術史学会という最大の学会においても、ＩＴ化の話題が出ることはほぼありません。そういう環境で、あらゆるフォーマット、或いは文字コード、種々雑多なものに検索をかけていく。

全文テキストとしては、まず、書籍から入れられたものがある。或いは私どもの場合のように、作品を調査してそのまま原資料からデータを入力することもある。それから、文献目録などでよくあるのは大型データベースなどからのCSV出力データです。そして、私もよく使いますが、Ｗｅｂページで後々使えそうなデータがあった場合それを保存しておく、というものです。それらのまさに雑多なデータを、今はハードディスクなども大容量化していますので、どんどん取ってきてためておくことができる時代になったのだと思います。

全文テキストデータの有用性

全文テキストデータの有用性というのは、大量のデータを早く、かつ正確に検索することに尽きると思いますが、それが利用可能な

時代になってきているということで、ますますその有効性というのは上がってきているのだと思います。検索が早いという点ではどんどん今後も期待できるわけですし、或いは検索の正確さということにおいても様々なノウハウが蓄積されてきつつある。

それから全文テキストの有効性のひとつに、これは非常に重要なことだと思うのですが、正に種々雑多なデータを扱えるということがあります。ですから、ある意味では、取り敢えず入力してしまえ、ということが可能である。そういう意味ではデータの作成が非常に簡易にできるということです。後は検索のスキルで何とかできる部分というのが、かなりあるだろう。それからデジタルデータですから置き場所に困らない。複製が簡単ですから皆で持つことができる。今の、インターネットの時代ですと、それこそＷｅｂ上に置いておけば誰でもそのデータを共有して利用することができる、という時代になっていると思います。まさに、そういう時代だからこそ全文テキストデータ、主として大容量のプレーンテキストデータベースの意義というものも増してきているのだろうと思います。

ＩＴ初学者による全文テキストデータ利用の問題点

データの入手方法

初学者の方にはこれから先の話が参考になればいいなと思っています。まず、ＩＴにあまり慣れていない人がどうやって全文テキストデータの利用に入っていくのかという点です。

第１につまずくのが、そのデータがいったいどこにあるのかということと、どうやってそのデータの在りかをつきとめるのかということだと思います。私もそうでした。たまたまZENBASEの場合は歴史系のBBSをのぞいていたために気がついた、というぐらいのことで、特に、例えば学生に使わせるといった場合にどうやるのかというのは非常に難しいところです。今のところ研究者であれば学会関係者との情報交換ですとか、いろいろなことでできてくるのだろうと思います。学生とか、初学者の方々には、インターネット上のデータの検索で引っかかる、サーチエンジンで引っかかるとかいうことが多いでしょう。或いはデータベースのディレクトリもＮＡＣＳＩＳなどで作ったりしています。そういうものを参考にして、きっとデータを集めることになるのだと思います。

データの保持

さて、ある程度大容量のデータが集まったとすると、そこで今度はそのデータをどうやって自分が持っておくのか、それをどう使うのかということが、おそらく問題になるのだと思います。持っておくためには、もちろんＣＤとか、Ｗｅｂ上でもいいのですが、最終的に使うのはおそらく手元のパソコンのハード・ディスクになると思うのです。サーバ・クライアント・システムがいかに発展しても、やはり私などの場合は調査先に持っていってそこで検索したいということがよくある。そういう場合には手元にパソコンが必要になる。パソコンが必要になるということは、当たり前のことですがパソコンが使えなければいけない。ところが、実際に学校で教えていても、学生ではほとんど大容量テキスト検索のノウハウ、スキルを持っていないというケースが普通なのです。それをクリアする必要がある。

データの性格分析

それから少し上級というのか、ただ単にGREPのソフトを使わせて検索しろというだけではなくて、もう少し有効な使い方を考えた場合に、師先生のお話でも「使う時には注意が必要です」と仰ったことのひとつが問題になります。それは例えば文字コードの問題であったりします。しかもまずいことに、例えばZENBASEやSATの場合というのは、こういうフォーマットでやっています、ということをきちんと宣言してありますから、使う側は非常にやりやすい。ところが昔作られたものなどはいい加減です。全然、宣言も何もない。とにかく入れました、が多いのです。それもやはり使わなければならない。その時に、自分でそのデータをある程度分析する必

要があるということです。

特にGREPソフトなどを使う場合に、当たり前のことですが、文字コードが違うとソフトで全くかからないということもありますし、違う文字がかかるということもあります。或いは入力フォーマット。これは1つは先程来も少し問題になっていた外字の扱い、外字をどういうタグで入れてあるかということ。それを知っていないと、外字は検索できない。或いは、意外と気がつきにくいのは改行、改ページはどういうふうに表現されているかということです。極端な例ですと、後ほどZENBASEで見て頂きますが、全く生のデータでは、単語が改行で切れてしまっているのです。そうすると改行を越えて検索する必要が出てくる。或いは逆に、何がしかの区切りが入れてある、例えば句読点が入れてある。アップ形式といって、ZENBASEはそういう形式でできているわけですが、そうするとその訓読、その読み方が間違いであった場合、マルが入っているために検索にかからない場合がでてくる。そういう問題が様々あって、ですから実際に全文検索する場合には入力されているデータのフォーマットであるとか、性格を十分に把握する、そうすることによって精度を上げる、或いは自分の目的の結果に近づくことができるわけです。これはどうしても必要なことだと思います。

検索方法

それからもっと具体的な話になってきますと、検索方法をどうするのか。おそらくごく普通の学生レベルでは、エディタやワープロに読みこんで、検索コマンドを使うということが行われていると思います。それでもよろしいのですが、通常、大規模データを扱うということになると、そのためのソフトを使うのが適当ではないかと思います。ただ、大規模な、例えば先程のＣＢＥＴＡであるとか、或いは検索系のサイト、大規模データベースによって検索をかけられているサイトなどは、汎用機を使って大きなインデックスを作って検索スピードを上げるということをしますが、なかなか個人には難しい。そのシステムを作るとなると機械語の勉強をしなくてはいけないなど、いろいろな問題が出てきます。では、そういった場合にはどうしたらいいのか、実は、今のパソコンのスピードですと、かなり大きなテキストでも、プレーンテキストであれば、それこそタダで手に入るＧＲＥＰのソフトで、相当なレベルの検索が可能になっています。こういうことを知っていると得です。

そのGREPをかけるソフトにしても対象のデータの性格、例えば大きさとか、異体字が統一されているのかどうか、というようなことまで含めてGREPのソフトを選んで、より効率的な検索ができるようにするのが適当だと思います。

それから、これも意外と大事なことかも知れないのですが、図書館員の方々はそう思われると思いますが、あるデータが置いてあります、で、難しいデータベースの利用方法だと、学生にはなかなかできないのです。その場合、結構インターフェイスが大事になってくる。学生でも使ってみようという気にさせる、それから学生でもある程度の検索結果を出せるようにするということが、それなりに重要になってくると思います。そのためには、それに適当なソフトウェアの選択が必要になります。もちろん市販の高価なものでもいいのですが、なかなか個人では買えない。そういう場合には、今では様々なフリーウェアとか、シェアウェアで、それなりにいいものがあります。

今日は3つほど、私が普段使っている、検索対象のテキストに合わせて使い分けているソフトを、後ほどZENBASEの検索を主体にした話で実際に見て頂こうと思います。

国際禅学研究所のページ

初めにZENBASE。国際禅学研究所のコンテンツですが、ＣＤ発行当時より新しいサイトに変わっています。ZENBASECD1リリース以降に作られたデータも、今はアップされています。もちろん、私はこの新しいものをダウンロードして使っているわけです。

単純なGREPソフト

まず、極めて単純なGREPソフト（SGREP）、

これはフリーウェアですが、単純ということは早いということです。単純とはどういうことかというと、これはシフトＪＩＳしか検索できないのです。一応アスキーの大文字、小文字の統一や区別はできます。それからサブディレクトリを検索することも可能です。では、私が普段使っている、美術史関係の文献データ、プレーンテキストで８メガくらいはあると思うのですが、これを「禅」というキーワードで検索にかけてみます。

いかがでしょう。意外と速いと思われたと思います。ただし、これは先ほど申しました通り、極めて単純であり、取り敢えず調べてみよう、という時に使います。あと、本当は郵便局のデータも大きいものですから、郵便番号を調べる時によく使います。

SFIND

それからインターフェイスという点で、学生などに使わせるのにはこれがいいのではないかと思うものを１つ。SFINDというソフト。これはシェアウェアです。

ZENBASEの中味は、ＣＤを開けますとフォルダがたくさんあり、HTMLが入っているものだとか、漢字ベースと言われる、先程師先生にご紹介頂いたものが入っているところとか、或いは文献データが入っているところとか、いろいろなツール類、文字コードの変換ツールとか、そういうものが入っているところがあります。その中の禅テキストと、後にリリースされたものを含めると、ＣＤリリース時には確か70幾つだったのが今では100個近くになりました。私の場合、主としてアップ形式といわれるものを中心にテキストにされたものをハードディスクのひとつのフォルダに入れてあるのです。このフォルダ全体に、「頂相」（お坊さんの肖像画、私の専門領域の１つなのですが、）で検索をかけてみます。【図版５Ａ参照】

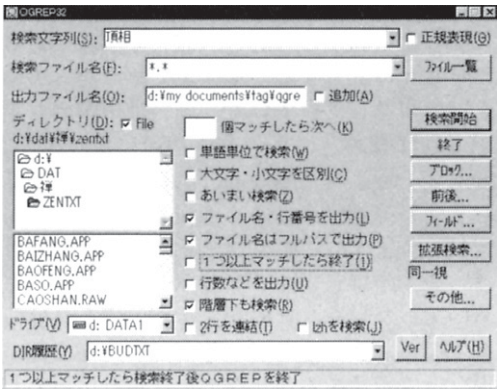
図版５Ａ「SFIND　キーワード『頂相』による検索結果」

このようにヒットした部分が上の段に出てきます。で、そこをクリックすると「タグジャンプ」をリアルタイムでやって、別画面に同時に表示してくれる。

これは全体を見ながら勉強していくのには非常に良いソフトではないかと思います。単純ですし、検索をかけるディレクトリを指定して、すべてのファイルを対象に、「頂相」という言葉で検索をかけます。ただしこれはＪＩＳとＥＵＣしか対応していません。且つ、細かい指定がほとんどできません。例えば正規表現とかそういうことはできないソフトです。でも、特にブラウジングしてというか、ある１つのキーワードで大量の文献を閲覧していくという場合には大変使いやすいソフトです。

QGREP

それから、もっともよく使うのが、と言うかほとんどこれしか使わないのですが、QGREPというフリーウェア。いろいろなサイト、全文テキストを載せているようなサイトでよく紹介されているソフトなのですが、これは極めて高機能です。正規表現にももちろん対応していますし、単語単位の検索、或いは大文字小文字の区別、あいまい検索というのもある。【図版５Ｂ参照】それから今こ



図版５Ｂ「QGREP　検索条件の設定画面」

こにチェックを入れているのは、いわゆるタグファイルへの出力命令です。検索結果から、タグジャンプという機能で元のデータに戻れるのですが、そのためのデータを出力しろという命令が出ている。或いは階層下のサブディレクトリまで検索するとか、それから、これも大変良い機能なのですが、２行連結というのがあります。先程少しお話しましたが、生のデータというのは、元の書物の改行で切ってあるケースが結構あります。実はZENBASEの中にもそういうものが入っています。RAW（ロウ）形式といいますが、そういうものを語彙検索したい場合はチェックを入れます。もちろんチェックを入れると時間はよけいにかかってしまうのですが、こういうものも初学者ー機械語を書いて、ということができない人間ーにとっては大変有用なものです。

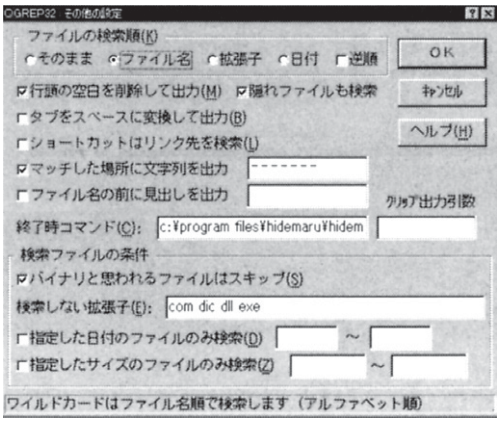
それからもう１つ、ここに今「同一視」と出ているのにお気づきになると思いますが、この同一視と同義語検索というのもあるのですが、それぞれ検索の目的に応じて使います。

同一視検索というのは、簡単に言うと、いわゆる異体字テーブルです。或いは全く違う字を同一のものと読ませることも可能です。先程台湾のものは「於」と「于」が一緒だと言う話がありましたが、これもその異体字テーブルにのせておけば、いっぺんに検索にかかるわけです。そのように自分でカスタマイズできるようになっています。また、自分で作った複数の異体字テーブルを切り換えて使えるというところが、またミソです。今現在私が使っている異体字テーブルはこういうものです。例えば仮名と片仮名を同じにするとか、日本の古典籍などの場合は仮名に濁音があったりなかったりするのですが、そういうものも同一視をして見ていく。或いは中国の文献であれば音通ということは普通にあります。日本でも禅宗の文献などには音通がよく出てきます。そういうものも１回で検索できるようになります。これはファイル名で区別されていますので、いくつかのパターンを作っておいて、プレーンな検索をしたい場合には数を少なくする、音通を主として引っかけたい場合には音通を主としたテーブルを作っておく、という形で対応が可能かと思います。それから、今は私はほとんど使っていませんが、同義語検索とは、シソーラスに対応するということです。この言葉とこの言葉を同じ意味と考えなさい、という形で検索をかけることが可能です。例えば大容量のデータベースで人名に検索をかけ、その結果をデータベースにして使うということがあります。この場合に、別名とか雅号とかいうものも同一視して検索するというような形でテーブルを作っておいて１度に検索をかける。これなどは正に初学者にとって、即そこである程度まとまった結果が出るという点からは、結構いい機能かなと思います。

検索結果の利用法

今、ＱＧＲＥＰというソフトを使ってZENBASEの全文テキストのディレクトリに検索をかけてみます。私の設定では、そのヒットした部分にこういうマークが入るように設定をしています。そうすると当たった部分が見つけやすい。【図版５Ｃ参照】

前後も読んでみたいという場合には、タグファイルという形式で、さっきのＧＲＥＰソフトが検索結果を出力してくれていますので、そこからタグジャンプをして元のデータにいく。【図版５Ｄ参照】タグファイルの内容は、フルパスのファイルネームと行番号です。元のデータに飛んでくれているのがお分かりですか？このようにして前後が読めます。タグジャンプをして元のデータの該当部



図版５Ｃ「QGREP 『その他の設定』画面」

図版５Ｄ「QGREP　検索結果から
タグジャンプで元のデータへ」

分に飛んでいけるわけです。そこで、自分の必要な部分を、ここまでは資料として使えそうだというところをコピーして、別ファイルに移してためておく。そうすると比較的簡単に自分で参照しながら勉強する資料集ができてしまう。

ただし、特に研究者の方の場合は、出てきた結果をプレゼンしなくてはいけないというケースがよくあります。たとえば、資料集を作って学会発表の時に配るとか、或いは教材としてＷｅｂサイトに掲示される場合も結構あります。そういう場合にはそこで表示できる形にしなければならない。これもそれなりのスキルがいるのですが。ここで最大の問題になるのは、外字です。表現できないと困る。このZENBASE、先程師先生にご指摘いただきましたが、CNSコードが使われているために、非常に手間なわけです。Ｗｅｂにアップするとか、刷り出すというのが非常に手間である。文字コード問題について詳しくない人間にとっては、外字を検索にかけることすら非常に厳しいということもあったりしますので、その辺は、例えば文字鏡なら、文字鏡には問題がありますが、比較的手元では表示しやすい。或いはGIFのフォントサーバのシステムもあります。CNSのコードを文字鏡のコードに誰かコンバートしてくれないかなというようなことを私などは思ったりしています。

特に学術利用の場合は、検索結果を原典にあたるというのは当然のこと、これは最初の沖本先生のお話でも出ていましたが、いろいろな意味でそれは当然のことだと思います。それで、ZENBASECD１の全文テキストはどうなっているかというと、先程の「宗鏡録」を、「頂相」で検索をかけると、大正新脩大蔵経48巻、661ページの３段目、18行目がヒットします。こういうタグと言うか、その位置を表すものが各行につけてあります。行単位で元に戻れるような配慮をしてある。これも、先程も申し上げたように２行にまたがる場合とか、いろいろなことがあって善し悪しなのですが、原典に戻るという意味ではこういう形式もあり得る。

或いは、実は国際禅学研究所のZENBASEでも使っているタブ形式というのがあり、これはページの先頭にだけページ番号が書いてあるというものですが、よく見かけます。これはまたそれなりに使いやすい場合があります。今見て頂いているのはアップ形式と言って、点マルがついている。このように切ってあり、どのように表現するかと言うと、例えば今のここを見ますと、これはＣの段の19行目という表示なのです。大正新脩大蔵経48巻の661ページのＣの段の19行目であって、この次の１は、前の行に１文字いっています、ということを表している。後ろを見ると、これは０ですから後ろはこの行でちょうど切れていますということです。その次の行で見てみますと、これは、前はそのままに切れていて、後ろの行に３文字送っていますということです。この形式では意味で行を切るため、原典の各行ごとでは切れないのでそういう表示をするようになっています。ですから、これを使えば簡単に原典に返って学術利用もしやすく作ってあるということです。

相当に超過しましたが、せっかくですのでZENBASEの自慢をもう１つだけさせてください。

ZENBASEの中に入っている大変便利なもの、文献データ、これは図書館員の方には非常に良いものかなと思います。いわゆるプレーンテキスト、或いはデータベースからはき出したCSVです。データベースの場合、容量が大きくなります、インデックスまで入れると極めて大きいわけですが、通常はこういうテキストで持っておけば十分GREPで検索で

きます。テキストで持てるということは、小さくて、且つテキストは単純なものですから、使いまわしが効くと、そういう例です。

それからもう１つは、禅語辞書データ。別にIMEのデータなどもあがっているのですが、それは別にして、このデータは大変便利なのです。そこに挙げております禅語辞書、禅学大辞典とか、非常に大きな辞書類、或いは解説類の見出しのデータを全て入れてあります。ですからこれで検索をかけて、あの辞書とこの辞書に出ているなと、で、そこのページにそのまますぐ行けるという、こんな非常に便利なものも、ZENBASEの中には入っています。

大幅に超過しましたので、これで終わります。何とも雑駁なお話で申しわけないのですがＩＴ初学者、禅宗初学者が使えば、ZENBASEというのはこういうものだということでお話をさせて頂きました。

（ふくしま　つねのり）