人文系データベース（哲学ＤＢ）の現状と開発
―― データベース入門以前

芦田宏直

The Present Condition and Development of Data Base in the Field of Humanities (Data Base of Philosophy)
— a pre-introduction to Data Base

HIRONAO ASHIDA

The consolidation of text data base in the humanities falls behind the times. Although each field of study strongly needs cooperative reserch, development operation on the conference level, or among conferences; the present condition is far from its intended purpose.

Data base in the humanities, having fixed users only, is not widely circulated at all. This situation has been caused mainly by two reasons.

First, there is no proper software for research. Secondly, data base users avoid intervention of publishing companies.

Here, I would like to think about several conditions of data base development, on the assumption that there will be wide circulation, as well as introducing the data base I myself have developed.

〔0〕はじめに

はじめにお断りしておきますが、私は、この情報知識学会に入会させていただいてまだ半年も経っていない言わば素人の会員だということ。しかも、私の専攻は、ドイツ・フランス近代哲学及び現代思想ですので、学会のみならず「情報知識学」にも全くの門外漢です。さらに今回の発表に際して多分野の会員の方がおられることを念頭においたこともあり、以下の私の発表のどの部分を取っても、聞いておられる方にとっては焦点の合わせづらい中途半端なものになっているかもしれません。この２点、あらかじめご容赦ください。

今回私たちが個人的な規模で開発したフルテキストデータベースは、すでに公刊されている哲学著作を MS-DOS テキストファイル化したものに独自に作成した検索プログラムを付加したものです。現在、クロスターマン社版ハイデガー全集の内、『存在と時間』『現象学の根本諸問題』『カントと形而上学の問題』『道標』『杣道』『哲学への寄与』『言葉への途中』『思惟とは何の謂いか』ができており、他に、ヘーゲルの『精神現象学』(ズールカンプ版)『大論理学』(同前)、カントの『実践理性批判』『判断力批判』『道徳形而上学の基礎付け』(すべてアカデミー版)、西田幾多郎『善の研究』(岩波文庫版)『西田幾多郎哲学論集全三巻』(同前)ができています。その他、ニーチェ、フッサール、ウィトゲンシュタインの諸著作の

若干が進行中です。特に、未公開の草稿、講義録等が続々と公刊されつつあるハイデガー全集のリアルタイムなデータベース化に、ここ２～３年の間に追いつきたいと思っています。

〔1・0〕 多数の、機械（コンピュータ）に無縁な研究者のためのデータベース

私たちがデータベースを作るときに考えたことは、〈誰にでも〉〈どこでも〉使えるものを作ることです。現在、人文系データベースは、先端の部分、たとえば、ＯＣＰや、マイクロＯＣＰのような「文書解析データベース」と、私自身が属している末端の部分、つまりワープロまではどうにかこうにか使えるようになったが、パソコンや MS-DOS のどうこうということになると皆目見当がつかない人達にとってのデータベースとの乖離が大きく、乖離以上に後者 ── 人文系の圧倒的多数の人達が属している ── にとってのデータベースは現在皆無といっていい状況です。これらの人達は大学や関係施設に高性能で高額なＯＣＲ（光学式文字読み取り装置）が導入されても、読み取ったテキストファイルをどう調理すればいいのかわからないまま、宝の持ち腐れ状態になっています。むろん、ＯＣＲという入力装置を知っている（ほどの）人であれば、それをエディタやグレプ類の検索機能で調理したり、FREE SOFT の TEXAS などを使いながら処理をすればいいことぐらい心得ているのかも知れません。しかし、私が知っている限りの末端の部分（大部分）の人達というのはそれ以前の人達であり、“ＯＣＲ”という言葉すら知らない人達のことを指します。もっとも、これらのことは必ずしも末端とは言えないのであって、たとえば、昨年末に出た『パソコン雑誌を読むためのキーワード』（朝日パソコン臨時増刊号）という最新の用語解説集では、ＯＣＲは、郵便番号読み取りか名刺読み取り程度の実用性しかないように「解説」されています(90ﾍﾟｰｼﾞ)。周知のように、現在、日本語ＯＣＲも読み取り率では欧語ＯＣＲに引けをとりません。私たちが使用したＯＣＲでも、たとえば西田の『善の研究』岩波文庫版で１頁あたり１～２箇所の誤読があるかないか程度のものでした。これは充分実用的です。この雑誌は、御存知のようにパソコン雑誌の中でも末端の、しかもその意味でも人文系の人が読む雑誌の代表的なものの一つなのですが、啓蒙誌がこの程度の認識であれば、その読者の現状は推して知るべしということになります。

〔1・1〕 データベース普及の問題

データベースが一部の機械（コンピュータ）好きの、あるいは機械（コンピュータ）好きの周辺にいる研究者にしか使われていない現状の問題点はどこにあるのでしょうか。

一つは、データベースを利用している（少数の）者と利用していない（多数の）者との落差から利用の恩恵が生じる場合があるため、つまり、データベースが普及していないことから利用の恩恵が生じる場合があるため、「引用」が外面的になり、読むに堪えない“データベース論文”が出回ることになるということ。今まで誰も指摘できなかったようなテキストの事実に驚くこと、場合によっては読んだことさえないテキストの事実に驚くことが先だってしまい、それに引き摺られ論文を内在的に構成することができない危険性が付き纏（まと）うことになります。もともと、未邦訳文献、未紹介論文の使用度合い、つまり新しい「引用文」の使用の度合いが、論文の評価をきめる場合も少なくはないこの分野の傾

向の中で、検索データベースが存在する、しかも、特定の者しか使えないデータベースが存在するというのは、「引用文」をますます外面的なものにしがちです。これは悪循環としか言いようのないものです。ことさらに引用される必要もない無用な引用文で綴られた“データベース論文”が、データベースを使ってもこの程度のものかというデータベース無用論を誘発し、そのためにふたたび使用者が固定化、孤立化し、再度「引用文」がインフレーションするという悪循環。

この“データベース現象”とでも言いうるものは、引用の装置としての検索データベースそのものの問題ではありません。研究者に限らずテキストに関心を持つ全ての人がデータベースを利用できる状態になれば、誰でもが自らで「新しい」引用文（テキスト）と出会う可能性をもつことになり、ことさらに「新しい」テキストを引用する必要もなくなります。もともと「引用文」とは情報の稀少性（情報の偏在）から生じていたものなのです。誰でもが知っている（知りうる）テキストをわざわざ括弧をつけて引用することはないのですから。

それゆえ、データベースは誰でもが使える環境で存在しなければ意味がない、つまり一気に存在しなければ意味がないと言えます。「自分が使える」だけではだめなのです。

私たちのデータベースは、MS–DOS テキストファイルを検索プログラムのためにいじる手間をなくすため、検索プログラムと一体化したデータベースにしています。そのため起動については、特別なインストールや前処理なしに SUZ（『存在と時間（Sein und Zeit）』の場合）＋リターンで立ち上がり、フロッピーに比べて処理速度を約10倍以上高速化できるEMSメモリーに自動対応します。操作については、ユーザーインターフェイスに配慮しマニュアル不要の対話型データベースにしたため、コマンド類を用いることなく、検索文字入力以外は全てファンクションキーを使うだけで済みます。パソコンをいわばワープロ専用機のようにしてデータベース専用機のように使うことができること。それが私たちがデータベースを普及させることの第一歩だったのです。

〔1・2〕頁テキストの表示

先ずインストールを含めて起動をワープロ専用機並みに簡単にすること。その次に私たちが考えたことは、書物のイメージを崩さずに検索できることです。現在、私たちのような末端の者に手に入る既存の検索ソフトは、グレプのような〈行単位〉の抜き出し型文字列検索ソフトにいくらかの変形を与えながら作られているか、それとも、大容量化したハードディスクの中で膨大化した諸テキストファイルを使用の目的に応じて捜し出すための〈ファイル単位〉中心の「テキスト型データベース」かどちらかです。そのため、両者とも、研究者や読者が書物の〈頁〉を捲りつつ言葉を捜し求めていくイメージからかけ離れていると言えます。９８（キューハチ）環境でフルテキストデータベースを使う場合には、特にこの傾向は強くなります。

検索結果を頁数や行数表示とともに前後何行かのテキスト表示にとどめるだけでは、ましてファイルの指示にとどめるだけでは、おそらく利用者は、再び実際の書物の頁を捲りつつ当該箇所を辿ることになるでしょう。或る種の二度手間を強いるのです。

愛着の或る書物の印象的な言葉を思い浮かべるとき、私たちはその言葉が存在する〈頁〉の或る箇所を空間的に思い浮かべるのであって、頁数や行数、ましてファイル名を

浮かべるわけではありません。研究者や読者が通常、机の上で書物に向かうイメージを損なうことなく検索できるためには、頁テキストの表示を中心としたデータベースの開発が必要なのです。IBM系のテキスト型データベースには、この種の使いやすいものが色々あるようです。しかし、９８（キューハチ）とIBM系との違いがどこにあるのかということ自体が私たち末端の者にとっては不明で、それらの上で“一太郎”をまともに動かすこと（インストールすること）さえも他人に頼んでいる現状ではIBM系というだけで、つまり日本国内ではさして普及もしていないソフト、しかもマニュアルが英語のソフトを使わざるを得ないこと自体、データベースを私たちから遠ざけることになっているのです。検索データベースがどんな多機能で高度な仕様を有していても、最低でも対話型のワープロ専用機並みに使えるものでなければ、普及は望み得ないのが実状です。マニュアルを読んだり、機械操作の習熟に時間をかけるくらいなら、専業のテキストを読んでいる方がましだと考えるのはもっともなことだからです。

さて、検索結果を絶えず書物の頁テキストとリンクさせて表示できるようにすること。たとえば、私たちのデータベースの単語検索モードでは、検索結果が先ず単語の出所箇所頁数とその頁における個数（頁数-個数画面）と共に表示されますが、出所頁数表示箇所にカーソルを移動し、そこでリターンキーを押すと、瞬時に当該頁テキストがモニターに現れ当該単語が黄色に色付けされて表示されます（白黒液晶モニタの場合は、反転表示）。エスケープキーで再び、頁数-個数画面に戻って任意の頁数表示箇所にカーソルを合わせて、リターンキーを押して、再び該当頁テキストを開くというふうに、絶えず頁テキストを参照しながら検索結果を確認していくことができるようになっています。一語句（概念）と他の語句（概念）との関連を問う複合（コンビネーション）検索、たとえば、「存在」と「時間」という語が同時に出てくる箇所の検索も、あくまで頁単位に、その組み合わせが同じ頁の中で出てくる場合（AND 1 キー：Ｆ１）、連続する２頁の中で出てくる場合（AND 2 キー：Ｆ２）、連続する３頁の中で出てくる場合（AND 3 キー：Ｆ３）の三つのモードで複合（コンビネーション）性の度合いを問うことができます。この検索結果も、先ず出所頁数が列挙され、任意の頁数表示箇所にカーソルを移動してリターンキーを押すと、その頁テキストが瞬時に開き、「存在」は黄色、「時間」は青色でというふうに頁テキスト内で色別に表示します。この場合も、エスケープキーを押してふたたび頁数表示画面に戻り、当該の頁テキストを呼びだし、というふうに頁数表示とテキスト画面を自由に行き来しながら、しかも頁テキスト内での単語の分布状況を色別に確認しつつ検索できるようになっています。

グレプ的な抜き出し検索の機能については、行単位という機械的な単位ではなく、該当単語を含む文頭からピリオドまでの〈センテンス〉を抜き出す機能（センテンスモード）を備えていますが、この場合も、抜き出された画面において F・9 に当てられた テキスト キーを押すと、頁数入力のウインドウが開き、任意の頁数字を入力すると頁テキストが当該単語の色付けと共に表示されるようになっています。代名詞が多い哲学の著作においては、センテンス、あるいは抜き出された行の前後のテキストが表示されない限り、意味を持たない場合が多いからです。単語検索の場合と同じようにエスケープキーで再びセンテンス抜き出し画面に戻り、再度任意の頁数字を入力し、というふうに、ここでも頁テキストと抜き出されたセンテンスを同時に確認しつつ検索できるようになっていま

す。

〔1･3〕グラフモード（語句の分布性）

インストールなり起動を専用ワープロ並みに簡単にすること。検索結果を頁テキスト表示と共に示すこと。その次に私たちが考えたことは、無機質な頁-行列挙を睨みながら、つまり無機質な数字を睨みながら、あるいは、機械的な行抜き出しを見ながら、そのつどテキストに戻りつつ目当ての箇所を捜し出すこととは別に、目的の語句の著作における分布状況をグラフ化して示すことです。

フルテキストデータベースが辞書や辞典類のインデックス型の検索と異なり、テキスト内の全ての文字に検索をかけることができるということの最大の意味は、単語の分布状況を示すことが可能になるということです。分布を示すということは、その単語が存在しない箇所を示すことができるということと同じことを意味します。これは、人文系、特に哲学著作においては重要なことで、言葉の浮沈は思想の浮沈と切り離すことができません。分布性は、一つの著作内ではその著作の構成性を浮き彫りにしますし、全集単位の分布状況で言えば、その思想家の「初期」から「後期」にかけての思想の変遷を問うヒントを与えてくれる場合があります。しかし、出所頁数や行数指示、あるいは行抜きだしというその語が存在する箇所の指示の形態だけでは分布性のイメージにかけると言わざるを得ません。私たちのデータベースは、単語検索モードの際、検索結果として出所頁数-個数を示した画面の［Ｆ・１］キーに［グラフ］モードを設け、そのキーを押すと検索結果を瞬時にグラフ化して表示することができるようになっています。画面左端縦列に一頁から順次頁が割り振られ、語句一つに付きアスタリスクマーク〔＊〕一個を当てる形で横に伸びる棒〔＊〕グラフ状のグラフが著作全体における当該語句の分布状況を一目でわかるようにしています。さらに、頁数と個数のグラフ化だけでは分布性の内容的な理解に欠けるため、［ＨＥＬＰ］キー（９８キーボード上、IBM系は［Ｆ・１０］）を押せば、著作の目次がウインドウ状に開き、構成的な理解の手助けになるようにしてあります。［ＨＥＬＰ］キーは、私たちのデータベースのどの画面でも著作の目次をウインドウ状に開く機能をもたせてあり、検索結果表示や頁テキスト表示の際の著作全体との関連を考慮する場合の手助けになるはずです。むろん、グラフ画面においても、気になる分布の頁テキストを見たければ、その該当頁を数字キーで入力してリターンキーを押せば瞬時に頁テキストが開き該当語句が色付けされて頁内での分布性も理解できるようになっています。頁テキストからエスケープキーで再度グラフ表示に戻ることもでき、分布表示から頁テキスト、頁テキストから分布表示という操作を繰り返すことで、従来のグレプ型の検索ソフトや“紙の”コンコーダンスとは異なるアプローチが可能になります。

〔1･4〕普及機（低価格機）使用と処理速度

この最低限必要な機能と操作の簡便性を、さらに普及機で処理速度を落とすことなく実現すること。現在、このデータベースは、PC-9800シリーズ(NEC）とIBM互換機で使えます。また現在、さらに高機能を付加したウインドウズ版を検討中です。データベース普及を前提にする私たちの最大の援軍は、何と言ってもIBM互換機の価格面での「大攻勢」で

した。現在、486 CPU（25MHz）搭載機で10万円前後まできているのは周知の通りで、高性能のハードが10万円を切れば、データベースのために（初めて）パソコンを個人（私費）で買うという人達が出てくるはずです。その上、大容量の記憶装置を必要とするフルテキストデータベースにとって、あるいはそれを高速度に処理しなければならないデータベースにとって、ハードの低価格化は必須の条件でした。また、いくら私たちのデータベース以上に多機能を誇るデータベースであっても一著作の単語検索に何十秒もかかるようでは実用的ではありません。そのうえそれが使いづらいものであれば、なおさら処理速度は気になるところです。論文を書きつつ、必要なフレーズの出所箇所が知りたい場合など、思考の流れを中断することなしに“待てる”時間は10秒以内でしょう。テキストファイルとして1.25メガバイトあるハイデガーの『存在と時間』で「世界（Welt）」という語を検索する場合、私たちのデータベースは 386（16MHz） EMSメモリー搭載機で約7秒、486（25MHz）EMSメモリー搭載機で約2.5秒で処理できます。フルテキスト型としては、何とか我慢できる速度だと思います。この速度が現在10万円前後のハードで実現するとすれば普及のためのハードの条件はとりあえず熟しつつあるのです。

（＊）以下、私たちのデータベース仕様の概略です。

★使用機　NEC PC–9800シリーズ、IBM–PC およびその互換機

★検索モード（検索範囲を一著作内の章単位、論文単位に指定可能）

単語検索モード（スペースを含む最大文字列７０字まで）

複合検索モード（スペースを含む一文字列最大７０字として最大２０組まで：AND/OR）

センテンスモード（スペースを含む最大文字列７０字までの語句を含むセンテンス抜出し）

★語句様態

検索語句の様態については、上記全ての検索モードでイタリック（ゲシュペルト）文字検索、前方一致／後方一致／部分一致、大文字・小文字の区別視／同一視、ハイフン（欧文右行末／語句内）の区別、行またがり文字列の処理など全てファンクションキー操作で可能

★印刷（PRINT OUT）機能

上記の全ての検索モードに用意されているファンクション 印刷 キー（F・1）で検索結果（グラフ表示を含めて）を印刷可能

★ページャー機能（目次を含めたフルテキストの頁行単位呼び出し、及びその印刷）

★ＥＭＳメモリー自動対応

〔2〕データベースの普及と商品化

こういったデータベース仕様の問題とは別に普及を妨げていることの問題に著作権の問題があります。現在、テキストデータベース利用は、“研究のため”という名目の下に出

版社との積極的な関係を避ける形で進んでいます。また、“研究のため”という研究者の誠意と熱意、場合によっては専門的な特定の関心が先行し、テキストファイル化から検索データベース作成まで研究者個人が担ったり、有志の研究者グループによる作成-利用の小さな輪の中で利用の特定化が起こりやすい現状にあると言えます。学会との関係でも、テキストファイル作成や検索データベース作成は研究成果それ自体とは厳密に区別されるため、正規の発表（公表）形式を取ることは不可能です。どんな道具を使ったかというよりも、それを使って何を書いたかということに学会が関心を持つのは明らかであるからです。現に私のここでの発表を私自身が一研究者として属する哲学の分野の業績発表と見なすのは難しいことで、またそれは当然のことだと言えます。私がデータベースの専門家でもないのに、この場に立っていること自体が、現在の人文系データベースの歪（いびつ）な状況を物語っていると言えます。そもそも、テキストファイル作成や検索データベース作成を人文系の研究者が担っていること自体が不自然なことなのです。これらの作業は、「この私」にしか書けない論文を書くという意味での論文業績の固有性なり、プライオリティーに比べれば、はるかに中性的なもので、取り立てて「この私」がやらなくてはできない仕事ではありません。にもかかわらず、仕事が完遂された場合には、誰が読んでいるか分からない紀要論文を書くよりははるかに寄与する率が高い仕事になるかもしれないという予感はあるわけです。それは、逆に言えば、自分の書く論文の諸々の特定性を越えて誰にでも役に立つ仕事をしているにもかかわらず、なぜ他でもない「この私」がデータベース作成に関わらなくてはならないのかという対照的な気分に晒されるということです。現在お茶だし事務職のようにＯＣＲを操作し、テキストファイル化を進めている院生や一部の若い研究者たちは、複雑な思いでこの作業に従事しているわけです。利用者が特定化し、データベース「利用の事実」や「成果」を報告しなければならないという疑似著作権的なルールでデータベース利用を私的に拘束せざるを得ないのは、“研究のため”の無報酬の“下部”研究者たちの屈折した感情が反映していると言えます。原テキストとしての書物の著作権以前に、データベース利用が内閉する構造があるのです。

この“問題”を回避するために私たちが考えたことは、データベース作成-開発は、版権を有する出版社自身の企画として進めるべきだということ。言いかえれば、データベースは商品化なしには、先の内閉する利用の特定化の環を打ち破ることができないだろうということです。

哲学の分野にとどまらず、現在著作物は、ほとんどの場合、印刷所にいく前にフロッピー入稿、つまりテキストファイルとして入稿されている状況で、書物の出版をテキストファイルとしても出版できる環境はすでに出版社の方で、つまり印刷所の手前で出来上がっている訳です。古典著作という前に、現在の著作物だけでもテキストファイル（テキストデータベース）として累積すれば、その利用の汎用性は格段に高まるに違いありません。このような出版にどんなリスクが存在するというのでしょうか。すでに著作権はコピー機によって事実上破壊されている現状で、テキストファイルが存在する分、少なくとも“読む”対象としての書物の売れ行きが左右されるとは思えません。また、販売の媒体は、紙の書物ではなく、CD-ROMやフロッピーであって、それらは書物の在庫を抱えてしまう危険性に比べれば、生産費コストの点でも、スペースコストの点でも比べ物にならないくらい効率的なものです。検索データベースということからすれば、当面需要は限られる（研究

者、大学研究機関、図書館など）のですから、注文販売の形態などをとれば危険負担は皆無と言えます。むしろ印刷所が印刷するための媒介変数のように使い捨てている（あるいは表（おもて）に出さない）テキストファイルを、また別の形態で商品にできるのですから、近頃はやりの“資源の有効利用”でもあるわけです。

通常著作権は、出版社と読者（利用者）とのいわば出口のところでばかり議論されていますが、生産-再生産の過程ではもはや「生産物（商品）」としての古典的「労働時間」（あるいは生産コスト）はほとんど関与していないといえます。また、引用の装置としてのデータベースによって、著者が剽窃され、著者自身も剽窃することになれば、もはや入り口のところでの「著者」「著作」という概念も稀薄なものになりつつあるのです。著作権の過度の主張は、反商品としてのテキストファイル（データベース）を出版社自身が有効利用する際の逆の足枷にもなりかねません。現在、事典、辞書類のインデックス型のデータベース化は出版社の方でも進みつつありますが、フルテキスト型の全集や単行本のデータベース化は出版社独自の企画としてはまだまだ、特に日本では遅れています。散在する研究者たちの特定利用の方が先行している状態だと言えます。今必要なのは、データベース利用を、学会以前の、有志研究者グループの有って無いような「交流」に任せるのではなく、それよりははるかに健全な（出版社の）商品流通にのせ、データベース普及を本格化させるべきだということです。データベース論文の、内容的にも形成過程としても歪な現状の次のステップを踏み出すためにも、出版社の認識転換を期待したいところです。私たちのデータベースは、MS-DOS テキストファイルになっているものであれば、ぼぼ一晩あれば、欧語・邦語を問わず全て先に紹介したものと同じ仕様で検索データベースに仕上げることが出来ます。

なお、私たちのデータベースは、私たちの趣旨も含め、岩波書店の好意的で積極的な協力を得て版権交渉も順調に進み、『存在と時間』（ニーマイヤー社版）、『精神現象学』（ズールカンプ社版）が近々同書店から出版されることになっています。

たしかに、私たちのデータベースは、本格的なパソコン使用による多機能で文書解析的なデータベースにくらべれば劣るかもしれませんが、商品（＝普及性）としての使い易さはとりあえず及第点を取れると思っています。なんだこんなものでも商品になるのかとお思いのパソコン使いの“プロ”の方の期待され予想される意見、またパソコン利用という点では末端の不慣れな者から“一太郎”を卒業した人まで各層に散在する研究者たちの意見を大規模かつ公開的に集約し、データベース普及と高機能化を健全な仕方で前進させるために、とりあえずこんなものでも商品として“形”にすることは必要だったのです。

以上、補うべきところは多々あるとは思いますが、これで私の発表を終わらせていただきます。

京都短期大学　芦田宏直
Kyoto Junior-College　Hironao Ashida