図書，論文，標本資料，映像音響資料，HRAF(人間関係地域ファイル)，などを収集し，研究に活用できるようコンピュータ化を行なっている．現在約130万件の民博独自のデータベースを構築している．これは単に文字表記による書誌的事項のみならず，できるだけ現物に近い情報を提供できるようなマルチメディアのデータベースを目指している．

標本資料の平面，正面，側面，鳥瞰からの映像の蓄積検索システムではすでに3万5千点分の映像データが取れている．1点の標本資料から7メガバイトのデータが発生するので，200ギガバイト以上が光ディスクに蓄積されている．スライドイメージの蓄積・検索システムでは35 mmのスライドをハイビジョンなみの分解能でデジタル化して入力し，各スライドに付けられた自然語の単語によって検索し，表示，プリントができる．また電子ファイルを使って本の頁イメージの検索なども行なっている．そのほか標本画像自動計測装置や，各種画像・音響情報の処理，地図と文化要素との重ね合せ(マッピング)，シミュレーションなど知的生産技術の開発などを行なっている．

人々の生活や自然環境の実態，文明の交流などを多面的に知るためには，単に映画やビデオだけではなく，文章による詳細な記述，写真，スライド，音など種々の情報媒体を総合的に用いる必要がある．しかもそれら異なるメディアの情報が連動して検索されることにより，理解が深まっていく．百科事典の説明と共に，資料の写真，それを使っている場面の映像，あるいはそこに生じる生活音や，現場における会話などが一体となることによって，あたかもフィールドにいるような臨場感が得られる．

このように静止画，動画などの映像情報，百科事典，民族誌などの記述情報，音声，音楽などの音響情報，それらすべてを収納し，検索し，相互に関係づけて提供するための仕掛け，これが民博の目指す総合的データベースである．

民族誌データベースとしてはテキストの全文を入力し，そのインデックスの自動作成や，ある事柄に関する記述がなされている個所を任意の単語で検索できるシステムも現在開発中である．すでに数百冊分のテキストをパンチャーによる手作業で入力しているが，文字自動読取装置を導入しテキスト入力の自動化も試みている．

これらいろいろな形態のデータを蓄積処理するためのコンピュータシステムとしてIBM3090・4341・9375，富士通M340R，VAX11/780・PDP11/60，TOSFILE，池上-RAMTEK，J-STAR，YHP，SONY-TEKTRONIX，その他，日本電気，APPLE，IBMなどのパソコン，画像・音響の入出力装置など各種のコンピュータをLANで結合したシステムを構成している．

著者推定問題の数理統計学的研究

統計数理研究所　村　上　征　勝

文献の数量的な性質，たとえば文献における単語の出現率，単語の長さの分布，文の長さの分布，品詞の出現率，延べ語数に対する異なり語数の比率などを調べ，その統計分析に基づいて文献の真偽判定や著者の推定などを行なう研究は，欧米では一世紀に近い歴史があるが，日本語文献に関しては最近ようやく本格的研究が開始されたというのが現状である．

本報告では文学作品，宗教書，哲学書，新聞記事，書簡などの著者推定を試みたいくつかの代表的研究と，用いられた種々の検定法，判別分析法，クラスター分析法などの統計手法を紹介し，それらの手法の適用妥当性について言及した．また現在進めている日蓮遺文の真偽判定

に関する研究を通じて得た，この種の研究を行なう上での次のような問題点

1 分析に用いるテキストに関する問題点
2 比較対照文献の選定における問題点
3 特徴抽出（分析に用いる統計量の選択）に関する問題点
4 日本語文献固有の分析上の問題点

についても報告した．

日本古代史の研究における一字索引の利用例

京都大学 計算機センター 星野 聰

筆者らは，既に六国史の一つである続日本紀のテキストを計算機可読形式とし，文献情報検索システムを用いて，オンライン検索を可能にしている．また，更に，日本古代史研究者の利用に便利なように，一文字ごとに用例検索ができる索引（以下において，一字索引という）を作成し，これを冊子体にしてプリントしたものを若干の日本古代史研究者の試用に供している．続日本紀のようなサイズのテキストに対しては，従来の手作業による索引作成は非現実的であり，計算機の利用が不可欠である．しかし，この種の開発には，現代の文書にない問題点が多く，専門的な知識が要求されるのである．

今回の研究発表は，このようなテキストに対する一字索引の，歴史研究に対しての有効な実例を示したものである．即ち，歴史研究上で一見意味のない，ありふれた文字，またはありふれた文字から成る文字列については，従来一般に作成され，利用されてきたいわゆる事項索引では，検索の手段がないのである．というのは，このような文字の重要性が認識されていないことが多く，事項索引では，恐らく索引のサイズを圧縮するために，見出し語として採用されていないことが多いためである．しかしながら，ありふれた文字でも，それを適切に解釈することが研究に大切な場合があるということを指摘した．

即ち報告では，上述のような文字の例として，「始」，「更」，「還」および「時」を取り上げ，また文字列の場合として「前年」という熟語について論じた．即ち，日本古代においては，「始」には単に開始を意味することがあること，「更」には変更を意味することがあること，「還」は元の場所に戻る場合に限って用いられること，「前年」は一昨年を意味することなどを実例を用いて示し，これらの解釈が歴史研究上で大切であることを説明した．これらの文字または文字列は，従来の事項索引では検索できないので，これに関する研究には一字索引の作成と利用が研究上で重要である．例えば，「前年」の語は，類聚国史索引にも類聚三代格索引にも収録されていない．また，考察の結果として，検索の対象とすべきテキストとしては，続日本紀以外の六国史，類聚国史，類聚三代格，万葉集などをも含むテキストを統合化したテキスト・データベースが必要となることが導かれる．また，一字索引を作成する際に，索引のサイズを減少させる目的で，もし，ありふれた文字を見出しとして採用しないとすると，研究の目的によっては，不便なことがあり得ることを示している．

歴史研究では，現存する限られた史料が意味する内容を，正しく理解しなければならないが，現代人の感覚で解釈したのでは，危険が多い．そこで，当時の関連する史料を抽出して，それに基づいた判断を行なうべきである．従来はこれを研究者の記憶に頼っていたので，見落としや忘却の危険があった．従って，計算機の利用は研究を能率的に進めることを可能にするだけでなく，研究のレベルを向上させるのにも役立つのである．