

研究者のためのマイ・データベース・システムの開発

及川 昭文

総合研究大学院大学 〒240-0193 神奈川県三浦郡葉山町
人間文化研究機構 〒105-0001 東京都港区虎ノ門 4-3-13

山元 啓史

オーストラリア国立大学

My Database System for Researchers

Akifumi Oikawa

The Graduate University for Advanced Studies: Shonan Village, Hayama-machi, Kanagawa
The National Institute for Humanities: 4-3-13 Toranomom, Minato-ku, Tokyo

Hilofumi Yamamoto

The Australian National University: Canberra ACT 0200, Australia

あらまし：一般的に人文系のデータベースは標準化が困難で、研究者それぞれの研究内容や成果と深く関連している。したがって、データベースの対象となる研究資料が同じであっても、データベースは研究者ごとに異なったものになりがちである。さまざまな研究資料がデータベースとして共有化されることは、その分野の研究の発展には重要な要素であるが、まずはそれぞれの研究者のニーズに応えたデータベースが作られることが必要となってくる。すなわち、人文系の研究者でもデータベースを簡単に作り、それを研究に活用できるシステムの開発が望まれる。マイ・データベース構想とそれを実現することを目標に開発したデータベース・システムについて報告する。

Summary : It is difficult to standardize databases in the humanities because the styles and formats of humanity database deeply depend on the contents and results of each study. It frequently occurs that the final format of databases by two researchers varies even if they use a completely same material. First, it is necessary to produce a database which satisfies the demands of researchers. Secondly it is essential to share a database with other researchers for the progress of studies.

We therefore propose a database management system called "My Database System (MDS)" which allows researchers to easily construct their own database and use it for their studies. We address the concept and the approach to the realization of MDS.

キーワード : マイ・データベース・システム, nihuONE, 品質管理, 研究支援ツール

Keywords : my database system, nihuONE, quality control, research tool

1. はじめに

マイ・データベース（以下、MyDB）構想は、筆者らが開発したBBDB（Bare Bone Database）システムにその緒がある。BBDBで目標としたことは、人文系研究者の中にある2つの技術的な課題の解決である。すなわち、まず、まず第1の課題は、資料をデータベース化できないということ。第2はデータ

ベース化までは何とかできるが、自ら使うのみで広くWebなどで公開はできないということである。BBDBでは“simple is best, small is beautiful”という理念のもとに開発を進め、1998年から総合研究大学院大学のサーバでの運用を開始した。その後バージョンアップや運用上の問題から、BBDBの開発の継続が困難になり、今日に至っている。[1]

そのような中で、大学共同利用機関法人「人間文

化研究機構（以下、機構）」によって、機構を構成する5つの研究機関（国立歴史民俗博物館、国文学研究資料館、国際日本文化研究センター、総合地球環境学研究所、国立民族学博物館）で保有するさまざまな研究資源を広く活用できる情報環境を創り出すことを目指して、3年計画で2005年度より研究資源共有化事業が開始された。

この情報環境を実現するために、3つのサブシステムに分けてソフトウェアの開発が2006年度から開始された。各研究機関が公開しているデータベースをダブリン・コアに準拠したメタデータに基づいて横断検索を実現するシステム（通称：分散型システム）、研究者自身によるデータベース作成を容易にし、また研究支援機能を強化したシステム（通称：集中型システム、以下、nihuONE）、時間情報や空間情報（地理情報）を分析できるシステム（通称：時空間システム）の3つである。[2]

nihuONEは、マイ・データベース・システム（以下、MDS）を強く意識して設計、開発されたもので、2007年4月から試行運用を始め、現在は第2次開発が進行中である。この第2次開発は2007年中に終了し、機能や性能についての評価を行い、その結果に

基づき最終的な手直しを行った後、2008年4月より本格的な運用を開始する予定である。

本稿ではMyDB構想を含め、このnihuONEでMDSがどのように実現されているかについて報告する。

2. MDSのコンセプト

データベースは基本的には多くの人に共有されることを前提として作成されるものであるが、MyDBとはその名のとおり自分だけのデータベースである。すなわち、研究者個人が、分析の対象として、あるいは研究の成果をまとめたものとして作られるものを想定している。

多くの場合、データベースは周到な準備を経て、組織的、計画的に作成されるものであるが、MyDBは研究者が個人的に（場合によってはグループもあり得るが）作成するものである。まずは作ってみて、不具合があれば項目の定義を変更したり、再編集したり、試行錯誤を繰り返しながら作られていくことになる。MDSは、そのような作業が容易に行われるよう設計されている必要がある。以下、MDSの要件について述べる。

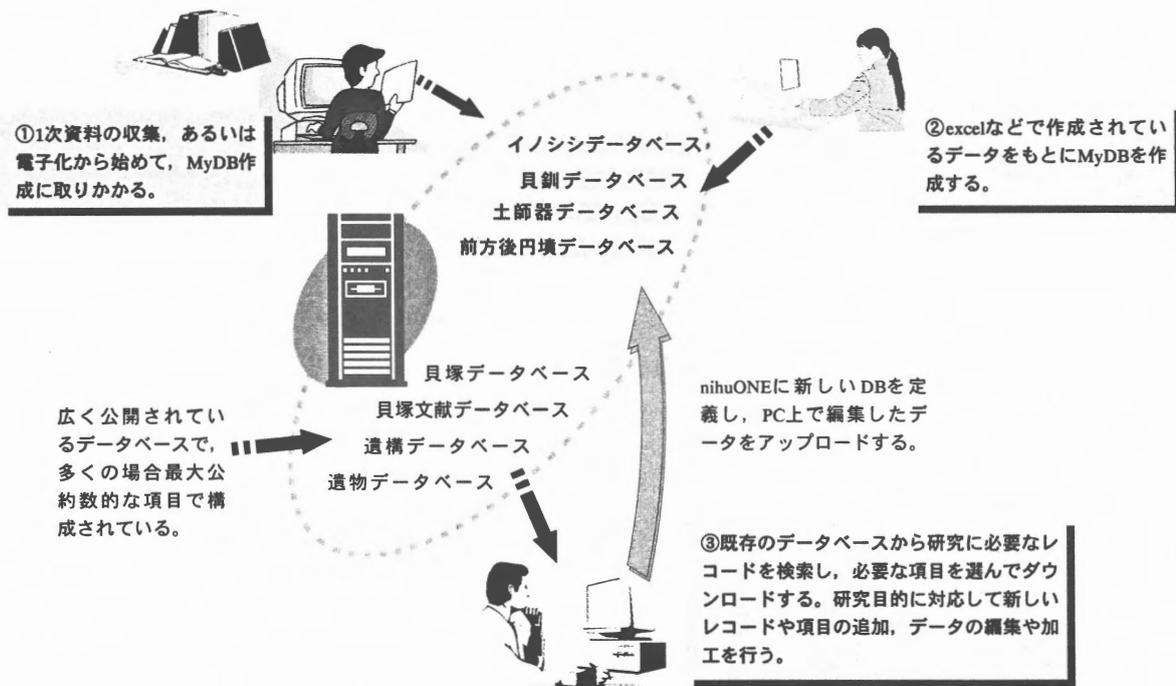


図1 マイ・データベース・システム (MDS) のイメージ

2.1 作るのが簡単

MDS の概念は図 1 で示されているが、MyDB を作成する過程としては、以下の 3 つのケースが想定される。

- ① 1 次資料を集めることから始める。あるいは 1 次資料は収集済みであるが、電子化はまだで、まったくゼロから MyDB を作成する (図 1 の①)
- ② 1 次資料の電子化はできており、excel などのソフトを利用してそれなりに活用している。このデータをもとに MyDB を作成する。(図 1 の②)
- ③ 既存のデータベースを加工、再編集したり、あるいは新しい資料を追加したりして MyDB を作成する。(図 1 の③)

いずれのケースでも、単純かつ容易な作業で MyDB を作成できることが、MDS には求められる。たとえば①の場合、excel などを使えなくても、ワープロなどで決められた書式でデータをテキストとして入力すれば、それがそのままデータベースにアップロードできること。②の場合では excel ファイルを CSV 形式のファイルに変更するだけでデータベースにアップロードできること。③の場合であれば、既存のデータベースから必要な項目を、CSV、タブ区切りなど多彩な形式で簡単にダウンロードできることなどである。

概要	鶏をモチーフとするコレクション
公開/非公開	<input checked="" type="radio"/> 公開 <input type="radio"/> 非公開 <input type="radio"/> 認証ユ
キーワード	鶏
ダウンロード	<input type="checkbox"/> 一般利用者可
URL	
一覧表示	<input type="radio"/> 通常 <input checked="" type="radio"/> サムネイル
備考	

図 2 データベース定義画面 (一部)

2.2 容易な定義変更

MyDB は研究と密着したデータベースとすることができる。したがって研究の進展に伴って、項目追加や項目の属性や内容の変更が必要となることが少なくない。また、研究成果をデータベースに反映し、そのデータベースを再び分析に活用するということもある。つまり、研究と連動してデータベースの構成や項目定義の変更を容易に行えることが MDS には不可欠である。

具体例で示す。図 3 は nihuONE で公開されている田村コレクション・データベースの検索結果一覧である。このデータベースは、すべてのレコードに画像が含まれており、検索結果一覧をサムネイル画像表示とすることも可能である。次期バージョンではこの切り替えを利用者サイドで、簡単に行えるよう

田村コレクション <検索結果 344 件>

次の 10 件>

No.	ID	枝	種別	素材	キーワード
1	10001	0	置物	プラスチック...	ページュ; オス
2	10002	0	置物	陶磁器	ミントン; グリーン; 茶; オス
3	10003	0	置物	ガラス	台座 (枠: 12*12**2 口); オス
4	10004	0	置物	陶磁器	
5	10005	0	置物	陶磁器	マイセン

図 3 検索結果一覧 -テキスト表示-

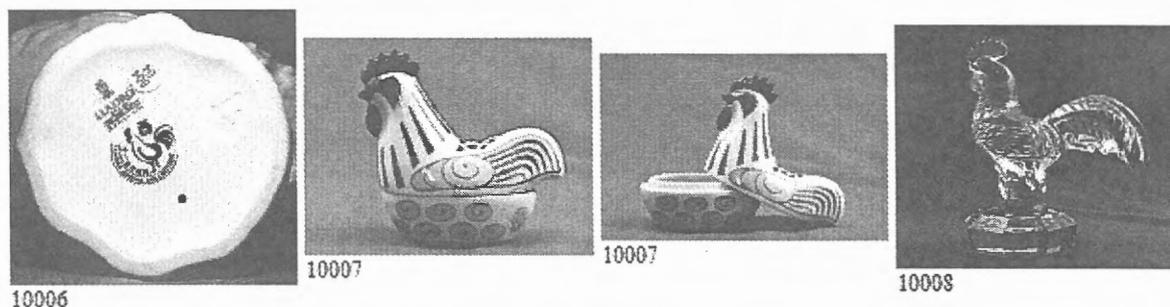


図 4 検索結果一覧 -サムネイル表示-

にする計画であるが、現在はデータベース定義で一覧表示を「サムネイル」に変更すれば（図 2）、データベースを作り直す必要もなく、即座に画像表示に変更することができる。（図 4）

2.3 検索から活用へ

本来データベースは蓄積されたデータを分析したり、人事システム、給与システムあるいは大学における学務システムなどのように業務を効率よく運用するために開発、発展してきたものである。ところが情報化時代の到来とともに大量の情報の中から必要なものを探し出す、いわゆる検索機能が重要視されるようになった。その結果、検索やそれに関連した機能は、ハードウェアや OS の発展とともに格段に充実してきているが、データベースに格納されているデータを分析したり、活用するための機能は不十分なままの状況にある。

MyDB は参照情報というより、分析対象として利用することを想定している。したがって、MDS では検索機能だけでなく、検索された結果の活用を支援する機能の充実が重要になってくる。どのような機能が必要かは、今後の大きな課題であるが、現在 nihuONE には支援機能としては、後述する統計機能、KWIC 検索機能とダウンロード機能などが準備されている。

3. nihuONE による MDS の実現

nihuONE の設計、開発においては、MDS のコンセプトを可能な限り反映させることを目指した。それがどこまで実現できたか、試行運用している現時点では的確に評価することは難しいが、前述した MDS の諸要件を nihuONE でどのように実現しているかを説明する。

3.1 nihuONE とは

図 5 のように nihuONE は XML 型データベース・エンジンである Shunsaku を核としているが、開発においては、まったく手を加えず製品版のまま利用することにした。これは Shunsaku をカスタマイズすることによって生じるシステム管理・運用の煩雑さやそのための経費増などを避けるのが大きな理由である。各モジュールの主な機能は以下ようになる。

システム管理・運用モジュール

- ・利用者の管理（登録・削除・編集など）
- ・データベースの登録・削除
- ・利用統計
- ・システムのバックアップとリカバリ

データベース管理モジュール

- ・データベースの管理
- ・項目定義
- ・表示（一覧、詳細）定義

検索モジュール

- ・検索（横断検索、詳細検索、絞込検索、数値検索、KWIC 検索など）

活用モジュール

- ・統計処理
- ・ダウンロード

各モジュールは、それぞれの改変やバージョンアップが他のモジュールに影響しないように、可能な限り独立性を保持するようにした。

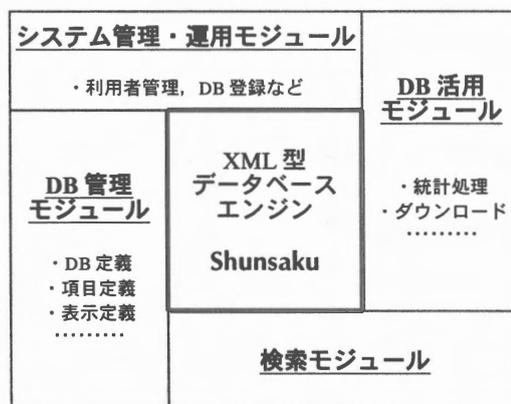


図 5 nihuONE の構成

3.2 システム運用と DB 運用の分離

一般的に DBMS の運用においては、システムの管理とデータベースの管理は同じ管理者が兼ねることが多い。このため DBMS に収録されるデータベースが多くなればなるほど、管理者の運用に関わる負荷は大きくなる。結果として、それぞれのデータベースのアップデートや保守に支障をきたすことが少くない。nihuONE ではこのような状況を回避するために、システム運用とデータベース運用を完全に分離している。

nihuONE の利用者は表 1 のように 4 種類に区分される。nihuONE での基本的なデータベース作成プロセスは、

- ①データベースの登録
- ②データベースの定義
- ③項目の定義
- ④表示（一覧，詳細）形式の定義
- ⑤データのアップロード

となる。このプロセスにおいて、admin が関与するのは、①のデータベースの登録のみである。

この登録処理では、「データベース識別記号」と「データベース管理者」の設定を行う。このように、①以後の作業はすべてデータベース管理者の仕事となり、システム管理者はデータベースの管理・運用の作業から解放されることになる。このことは多数のデータベースを公開、運用している部署にとっては、システム運用にのみ専念すればよく、非常に大きなメリットとなる。

一方データベース管理者は、自分が作成し、その内容を熟知しているデータベースのみを管理するわけであるから、たとえ複数のデータベースを管理するとしても、運用の作業は大きな負担とはならない。むしろ、システム管理者に依頼しなくても、自分の判断でそれぞれのデータベースを管理・運用できることは大きな利点といえることができる。

表 1 nihuONE の利用者区分

区分	権限・役割など
admin	システム管理者で、一人だけ登録できる。その役割は利用者の管理、データベースの登録・削除、利用統計の閲覧、システムのバックアップ・リカバリなどである。
DB 管理者	データベースの管理・運用を行う者で、各データベースに必ず一人だけ設定する。
登録利用者	認証を必要とするデータベースの閲覧やデータのアップロード・ダウンロードの権限を付与された利用者である。
一般利用者	一般公開されているデータベースのみを閲覧できる利用者である。

3.3 項目定義と品質管理

データベース管理者がデータベースを作成するにあたって、最も重要な仕事は項目定義である。項目定義では表 2 で示す属性を各項目ごとに設定することになる。

この項目定義は画面（図 6）上で行うが、あらか

じめ excel などで作成しておき一括してアップロードすることもできる。また定義が終了した時点で CSV ファイルとしてダウンロードする機能もある。

項目定義は、既に入力されているデータの内容との整合性が保証されれば、データベース作成後に変更することもできる。たとえば、検索項目となっていない項目をあらたに検索対象としたり、区切り文字を追加したり、最小値や最大値の値の変更などである。また、項目そのものを新しく追加したり、削除することも可能である。

データベースの利用価値を左右するものの一つはデータの品質である。nihuONE においては、データの品質を高めるための品質管理（Quality Control, 以下 QC）機能を充実させた。QC は項目定義でパラメータを入力しておくことで、データのアップロード時に実行される。表 2 は項目定義で設定しなければならない属性の一覧であるが、そのうち属性の欄が塗りつぶされているものは QC 項目である。

データ型チェック：その項目が表 2 の「データ型」のどれにあたるかは必ず指定しなければならない。それぞれの項目データがこの指定に適合しているかどうかを調べる。

ペアチェック：文献目録データベースなどで「著者名」「著者名よみ」のように必ず同時に存在しなければならない項目があった場合、いずれかが欠落していないかを調べる。

Table 変換：入力されたデータが、DB 管理者によってあらかじめ登録されているテーブルに含まれているかどうかを調べる。テーブルには変換する値（数値でもテキストでも可）も設定されており、エラーがなければその値を指定された項目のデータとして展開する。

デリミッタ項目の 2 重チェック：たとえば「シジミ、アサリ、ハマグリ、カキ、アサリ」と入力されたデータを、「アサリ、~~アサリ~~、カキ、シジミ、ハマグリ」とソーティングし、同じ文字列がある場合は自動的に削除する。

辞書チェック，辞書変換：あらかじめ nihuONE にデータベースとして登録されている辞書と照合し、入力されたデータがその辞書に含まれているかどうかを調べる。また、エラーがなければ、その辞書の任意の項目値を新しい項目データとして展開する機能もある。

B01 貝塚データベース

項目自動生成 項目数: start: 間隔:

項目定義アップロード ファイル名:

項目定義ダウンロード CSV形式

タグ	項目名	ID項目	データ型	種別	必須	非公開	htmlテキスト	検索	KWIC	統計	頻度	複数項目区切り	ソート	削除
編集	010 遺跡番号	<input checked="" type="checkbox"/>	11	D	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	削除
編集	020 県市コード	<input type="checkbox"/>	2	D	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	削除
編集	030 遺跡名称	<input type="checkbox"/>	3	D	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	削除

図6 項目定義画面 (部分)

表2 項目定義 -属性一覧-

属性	内容	属性	内容
タグ番号	項目を識別する3桁の数字	KWIC検索	KWIC検索の対象とするかどうかの指定 (Yes/No)
名称	項目名称	区切り文字	デリミッタ項目 (注2) の場合, 使用する区切り文字。
表示名称	一覧表示や詳細表示の際に表示する項目名称	ソート	デリミッタ項目の場合, 区切られた文字列単位でのソーティングの指定 (注3)
ID	ID項目かどうかの指定 (Yes/No) (注1)	html	htmlテキストかどうかの指定 (Yes/No)
データ型	データの種別を表すコード 1=テキスト (全角のみ) 2=テキスト (半角のみ) 3=テキスト (全角, 半角) 11=数値 (カンマ, 小数点を含まない) 12=数値 (カンマを含む) 13=数値 (小数点を含む) 14=数値 (カンマ, 小数点を含む) 20=ファイル	頻度統計	検索結果の集合に対して頻度統計の対象とするかどうかの指定 (Yes/No)
最小値	データ型が数値の場合の最小値を指定 (不明の場合は空白)	基本統計	検索結果の集合に対して基本統計の対象とするかどうかの指定 (Yes/No)
最大値	コードが数値の場合の最大値を指定 (不明の場合は空白)	Table変換	Tableを利用したテキスト変換の指定
テキスト長	コードがテキストの場合の最大バイト長を指定 (不明の場合は空白)	リンク対象	他のデータベースへのリンク指定
必須項目	必須項目かどうかの指定 (Yes/No)	ペアチェック	著者名と著者名読みのように必ず同時に存在する項目の指定
検索対象	検索の対象とするかどうかの指定 (Yes/No)	辞書チェック	辞書を参照してその辞書の中に存在するかどうかをチェックする
		公開・非公開	公開・非公開の指定
		ダウンロード	ダウンロードの可否の指定

注1: 複数の項目を ID 項目として指定することもできる。その場合すべての項目値を連結したものが ID として認識される。

注2: 「デリミッタ項目」とは、一つの項目に複数のキーワードなどがあらかじめ指定された区切り文字 (デリミッタ) で区切って入力されている項目のことである。たとえば、区切り文字「;」で区切られたテキスト「情報処理; データベース; XML」などが入力されている項目。

注3: デリミッタ項目の場合、区切り文字で区切られた文字列単位でソーティングして格納しなおすかどうかを指定する。例えば、「アサリ, シジミ, マシジミ, ハマグリ」と入力されていた場合、これを「アサリ, シジミ, ハマグリ, マシジミ」とソーティングすることである。

エラーログ：QC 機能ではないが、エラーログにも工夫を加えた。図 7 はアップロードの際のログリストである。ここではいくつかのエラーがあることが表示されている。エラーが生じた箇所を見つけやすいように、正しくアップロードできた最前のレコードの ID を表示し、そこから何行目にエラーがあったのかが示されている。すなわちこのレコード ID を手がかりにすれば、もとデータのエラー部分を探すことが容易になる。

```

Batch START
クラス名: com.fujitsu.campus.nomir.control.batch.FileUploadBatch
ユーザID: admin
開始時間: 2007 11 07 02:30:59
貝属性データベース のアップロードを開始しました。
27行目 (10025から1行目) 040: 水深11には数字を入力して下さい。
45行目 (10043から1行目) 040: 水深11には数字を入力して下さい。
46行目 (10043から2行目) 040: 水深11には数字を入力して下さい。
47行目 (10043から3行目) 040: 水深11には数字を入力して下さい。

```

図 7 エラーログの例

3.4 SDF 形式

レコード・データを一括してアップロードするファイルの形式としては、CSV、タブ区切りなどが一般的であるが、nihuONE ではこれらの形式の他に独自の形式として SDF (Standard Data Format) と名付けた形式を取り扱えるようにした。

SDF とは図 8 のように「タグ番号」「|」（区切り文字として使用）」とデータから構成されたテキストファイルである。したがって、ワープロなどで作成したり、編集することが可能となる。この SDF 形式を設定した主な理由は、CSV、タブ区切りの形式ではテキスト中に「改行」の制御文字を含めることができないからである。

たとえば、最近文献目録データベースなどに「目次」という項目を設定し（図 8 参照）、そこに目次の見出しを「/」などの文字で区切って入力する例が多く見られる。しかし、見出しが多数ある場合は非常に読みにくくなる。この「/」の代わりに「改行」にすれば、図 9 のように見やすくなる。

また、一つの項目のデータが長い場合（たとえば、貝塚データベースの出土している貝類の名称を入力する項目で、非常に多くの貝が出土している場合、テキスト長が 3000 バイトを超えるレコードがある）、excel の一つのセルで画面がいっぱいになり、編集がやりにくいということも理由の一つである。

```

010|10001
020|鉦野県発見『父母恩重経』の『省略経』の新資料（鉦
7）について
040|新井 慧誉
070|豊山教学大会紀要
090|第 30 号
100|13 東京
110|豊山教学振興会
120|2002/11/1
130|p17-28
150|石碑の概要／石碑建立のいきさつ／（鉦 7）のテキ
スト／経文の検討／結語
160|630
170|690
210|0388-2500
;

```

図 8 SDF 形式の例

13 ページ	p17-28
14 一般注記	石碑の概要 石碑建立のいきさつ （鉦7）のテキスト 経文の検討 結語
15 内容	

図 9 テキスト中に改行を挿入した例

3.5 リンク機能

データベースに「参考文献」といった項目がある場合、そこには「書名、著者名、発行者、発行年月日」といった書誌情報が入力されている必要があるが、階層構造を持たせて入力している DBMS も少なくない。nihuONE においても階層構造を持ったレコードを構築することも可能であったが、BBDB の理念である“simple is best, small is beautiful”を継承し、レコードの構造は可能な限り単純な構造とした。

フラットな構造にしたわけであるが、階層的な関係を実現するために「リンク機能」というものを実装した。この機能は次のような手順で実現する。

- 1) 項目定義の際に、あらかじめリンクするデータベースの識別記号を設定する。この項目を「リンク項目」と呼ぶ。設定できるデータベースは nihuONE に登録されていればどのデータベースでも可能である。すなわち同一データベース中の他のレコードにリンクすることもできる。
- 2) リンク項目にリンク対象となるデータベースのレコード ID を入力しておく。
- 3) 詳細表画面で、リンク項目に表示されている ID をクリックすれば、当該レコードが別画面で表示される。

7 時代	縄文中期
8 遺跡コード	113
9 遺跡種別	貝塚
10 遺構コード	999
11 遺構種別	その他
12 土器型式	早稲田3類・赤御堂
13 関連文献	020001; 020011; 020012
14 貝類	アカニシ, アサリ, ア ウバガイ, ウミミナ, タマガイ, サルボウ, ヤマトシジミ

データベース名		貝塚文献
No.	項目名称	入力内容
1	文献番号	020011
2	タイトル	赤御堂遺跡発掘調査概要報告書
3	著者	滝沢幸長; 松山力; 小滝一三; 崎田健; 工藤尚克;
4	掲載誌等	
5	巻号	
6	頁	29-33
7	発行年	1975
8	発行者	八戸市教育委員会
9	コメント	6-F 第一地点出土の貝類について: 小滝一三

図 10 リンク項目の例 (貝塚データベースと貝塚文献データベース)

具体的に示すと、図 10 は貝塚データベースの検索結果の詳細表示であるが、「関連文献」がリンク項目となっており、ここにある「貝塚文献データベース」のレコード ID (たとえば 020011) をクリックすると当該文献が表示される。

3.6 ファイル型項目の活用

項目属性のデータ型に「ファイル」がある (表 2 参照)。この項目に入力されるのはファイル名で、ファイル本体は事前に nihuONE のサーバ上に転送されることになる。詳細表示画面でこのファイル名をクリックすると、ブラウザがファイルの拡張子に応じた処理を実行する。たとえば、jpg, gif などの画像ファイルや pdf ファイルなどは、画面上に展開、表示されるが、CSV などのファイルは「開く」か「保存」を問い合わせる。

図 11 は幕末明治地勢地図の詳細表示で項目「索引地図」に入力されている pdf ファイル名をクリックすると図 12 のような地図が表示される。利用者はこの地図をダウンロードして自由に利用することができる。[3]

これらのファイルが nihuONE 以外のサーバに格納されている場合は、別の方法で同じような処理を実現できる。項目の属性を“html”と指定することによって、その項目に入力された文字列は html テキストとして解釈される。たとえば、 [のように入力しておけば、nihuONE 以外のサーバに格納されている画像が表示される。この機能を利用することによって、画像ファイルだけでなく、さまざまなファイルへのアクセスが可能となる。](http://www.minpaku.ac.jp/marsimages/mcd5/083.jpg)

項目名称	入力内容
国ID	K01
国名	山城
国名よみ	やましろ
郡数	8
郡ID	G01001,G01002,G01003,
郡名	葛野, 愛宕, 乙訓, 紀伊
郡名よみ	かどの, をたぎ, おとく
隣接国数	6
隣接国ID	K02,K03,K05,K06,K21,
隣接国名	大和, 河内, 摂津, 伊予
府県ID	M26
府県名	京都府
師管ID	S04
師管名	第4師管
索引地図	K01-1.pdf
pdf_files	K01-1.pdf; K01-2.pdf;
SHP_files	K01; G01
備考	

図 11 幕末明治地勢地図の詳細表示



図 12 pdf による地図

- 14 布地特性 刺繍, キルティング, 刺し
- 15 素材コード E20
- 16 素材 毛
- 17 構造技術コード F47;F49;F59;F76
- 18 構造技術 袖がついているバスト, 田中千代コレクション 黒
- 19 備考 女性用 上衣
- 20 標本名 田中千代コレクション
- 21 収蔵場所



22 サムネイル画像

図 13 国立民族学博物館にある画像の表示例

図 13 はその例であり、ここに表示されている画像は国立民族学博物館のサーバに格納されているものである。[4]

4. MyDB の活用

データベースを単に検索するだけでなく、そこに蓄積されているデータ(情報)を活用するためには、そのための仕組みをデータベース・システム側に準備しておく必要がある。nihuONE は“探すから使うへ”ということを開発の目標の一つとしており、そのための機能の充実を図った。具体的には統計処理機能、KWIC 検索機能の実装と、ダウンロード機能の拡充である。

4.1 統計処理

検索して得られたレコード集合に対して統計処理

を実行する機能で、「基本統計」と「頻度集計」の 2 つがある。

基本統計 データ型が「数値」の項目に対して「最小値、最大値、平均値、標準偏差値」を計算する機能である。

頻度集計 項目にデータとして入力された文字列単位でその頻度を集計する機能で、デリミッタ項目の場合にとくに有効である。2 つの項目までのクロス集計も可能で、さまざまな分析の基礎資料として活用できる。

基本統計もそうであるが、この頻度集計の結果は CSV 形式でダウンロードできる。図 14 は貝塚データベースの出土遺物である「哺乳類」の集計結果をダウンロードし、excel で若干の編集を行い表示した例である。

	A	B	C	D
1	ほ乳類	出現頻度	比率(%)	累積
2	イノシシ	1188	13.53	13.53
3	シカ	1080	12.3	25.82
4	イヌ	601	6.84	32.67
5	ウマ	495	5.64	38.3
6	ニホンジカ	448	5.1	43.4
7	タヌキ	385	4.38	47.79
8	ウシ	232	2.64	50.43
9	ノウサギ	195	2.22	52.65
10	有	193	2.2	54.84
11	アナグマ	186	2.12	56.96
12	クジラ類	185	2.11	59.07
13	クジラ	170	1.94	61
14	キツネ	127	1.45	62.45

図 14 頻度集計結果を excel で表示

4.2 KWIC 検索

KWIC (Keyword In Context) リスト(図 15)は、言葉の使われ方などを分析するのに利用されるが、ほとんどの DBMS には KWIC 検索は実装されていない。nihuONE では検索のための索引作成が不要であるという特長を生かし、KWIC 検索の実現を図った。

- パネルディスカッション「パソコンソフトに求められるもの - 人文科学者からの提案 -」
- メタデータによるデータベースの機関間連携の実現 - 人文科学 データ共有のための標準化 -
- 「人文科学とコンピュータ研究会」2年間の主査を終わって
- パネル討論「人文科学 にとっての“デジタルアーカイブ”」
- パネル討論「人文科学 研究資源の共有化 -国文学研究資料館を例として-」
- ...lin CoreメタデータとZ39.50プロトコルにもとづく 人文科学 系データベースの統合検索に関する実証実験

図 15 KWIC 検索結果

4.3 ダウンロード機能の拡充

利用者が多くなれば多くなるほど、データベースを活用するための機能は多種多様なものが必要になってくる。しかし、開発や運用コストを考慮すると、それらのニーズのすべてに応えることは困難である。nihuONE では次善の策として、検索処理のさまざまなフェーズで、データのダウンロードを可能にすることによって、利用者の利便性を高めることにした。すなわち、利用者は自分のパソコンにダウンロードすることによって、データを自由に加工したり、より高度な統計分析を行ったり、あるいは GIS などの入力データとして活用することができる。利用者がダウンロードできるのは、

- ①検索したレコード群の全データ
- ②検索結果一覧
- ③ KWIC リスト
- ④基本統計結果
- ⑤頻度集計結果

で、①の場合は、ダウンロードする項目やデータ形式の選択が可能となっている。②～⑤は CSV 形式のみとなっている。

5. 今後の課題

MyDB というコンセプトが、またそれを実現することを目的の一つとして開発された nihuONE が利用者に受け入れられるかどうかは、まだ不明である。現在機構の研究資源共有化事業のなかで、モニターを募集し、実際にデータベース作成が進められており、それぞれのデータベースが完成した時点で、これらのモニターに nihuONE に対する評価を依頼することになっている。その時点で要望あるいは課題が出てくると想定しているが、これまでの試験運用のなかで検討を要する課題として、以下を挙げることができる。

グループによる MyDB 作成 MyDB は一人で作成することを前提としているが、グループで一つのデータベースを作成することに対応した仕組みの検討が重要であると考えている。たとえば、全国規模の考古学遺跡データベース作成にあたって、日本全国をいくつかの地域にわけ、それぞれの地域を分担して複数の利用者がデータ作成を行う場合、その手順や品質管理をどのようにして行うかなど、検討すべ

き項目は多い。

データベースの特性分析 データベースそのもののバージョンアップ、すなわち項目の再編成やキーワードの変更などを検討するには、データベースがどのような項目や検索語で検索されているのかなどについての詳細な分析が必要となってくる。そのため、当該データベースの利用ログを一定期間収集し、基礎資料の作成を行わなければならない。現在の nihuONE にはその機能は実装されていないため、できるだけ早い段階でその機能の追加を実施したい。

6. おわりに

MDS は特定のデータベース・システムを指すのではなく概念的なものである。MDS にどのような機能が必要で、どのように運用されるべきかといったことは、まだ定まっているわけではない。MDS 構想そのものを含め、今後 nihuONE 利用者を交えて広く議論していくことを考えている。本稿がそのためのきっかけになれば幸いである。

- [1] 及川昭文, 山元啓史「Web 公開のためのデータベース・エンジニアリング」情報処理学会研究報告 CH-49, pp.49-56, 2001
- [2] 及川昭文, 藤沢桜子, 洪政国, 山元啓史「研究支援機能を強化したデータベース・システムの開発」第 13 回公開シンポジウム「人文科学とデータベース」論文集, 2007 (印刷中)
- [3] 鎌田聖子, 及川昭文「幕末明治地勢地図境界データの作成」人文科学とコンピュータシンポジウム論文集, 2007 (印刷中)
- [4] 中川隆, 高橋晴子, 及川昭文「民博コスチュームデータベース (MCD) の過去・現在・未来」人文科学とコンピュータシンポジウム論文集, 2007 (印刷中)