古今和歌集パラレルデータベースと公開システム*

山元 啓史† 及川 昭文‡

†カリフォルニア大学サンディエゴ校
‡総合研究大学院大学

要 旨

数理解析を目的として，古今和歌集 (以下，古今集) データベースとそれらを編集，公開，集計するためのマネージメントシステムを開発した．データベースは標準表記に従う 1,111 の和歌データのほか，英翻訳データベース，作者データベース，品詞解析データも作成し，それぞれ並行に登録し，検索，計算できるようにした．データベースマネージメントシステムは，BBDB というデータベース公開システムと BBQC という品質管理システムで，構成される．このシステムを利用して，上記古今集 DB の要素を構築し，公開，検索だけでなく，著者毎，和歌毎，品詞毎，等の集計が行えるようにした． 本論文では，古今集の数理解析の予備段階として，1) データベースの構築方法，2) マネージメントシステムの設計と仕様，および，3) データベースの利用の実際，について述べる．

Kokin Waka Shū Pararell Database and Management Systems

Hilofumi Yamamoto† Akifumi Oikawa‡

†University of California, San Diego
‡Graduate University for Advanced Studies

Abstract

We have developed the Kokin Waka Shū database (Kokinshū), the database of the collection of Japanese classic poems by Imperial order, and database management system in order to analyze Japanese classic poems. The database contains not only the general information of 1,111 Japanese classic poems in the original, but also the translations in English, and the parts of speech of each word in both Japanese and English. The database management system consists of two components: a database publishing system called "Bare Bone Database (BBDB)" and a database quality control system called "Bare Bone Quality Control (BBQC)." Using this management system, all the elements of the Kokinshū DB have been combined systematically, and users can not only search the information they want, but also calculate the number of authors, poems, words by the parts of speech, and so forth.

In this paper, a preliminary report on the Kokinshū project describes 1) The process of building a database, including the selection of categories; 2) The design and the development of management systems; and 3) The use of pararell Kokinshū database.

*本研究は，文部科学省科学研究費補助金特定領域研究 (A)118「古典学の再構築」の助成を得た．

1 はじめに

本論の目的は，古今集和歌集の英語翻訳を利用した日英パラレルデータベースの開発とそのマネージメントシステムの開発について報告することである．

前者であるが，このような古典文学の日英対訳のデータベースは，一般的に海外の日本文学研究者あるいは海外で行われている日本研究とりわけ，和歌の教育に役立つものと思われる[1]が，筆者らは特に，古今集という共通の材料を通して，外国人のもつ日本文化観が日本人のそれとどのように違うか，比較分析することを第一の目的とし，その基礎資料としてのデータベース開発を進めている．

ところが，データベースの開発を進める過程で古今集収録の和歌だけでなく，技巧のデータ，品詞データ，作者データ，複数からなる翻訳データ，統計分析のための前処理データなどが発生し，それらをまとめて管理する仕組みが必要となった．

そこで，後者の目的であるが，管理だけでなく単独あるいは複数を同時に公開，検索，計算するためのシステムの設計と開発を行った．設計には，筆者らが使うだけでなく，さまざまな，場合によっては，情報処理の技能・知識を必ずしも有していないユーザによる利用やデータベースの共同開発の可能性も考慮した．

以下において，コンテンツ開発として古今集データベースの開発，システム開発としてマネージメントシステムの開発を述べた上で，システムの利用の概略を説明する．

2 古今集データベースの開発

和歌を中心とする古典文学領域のデータベース化はかなり活発であり [1, 2]，国内だけでなく，海外においても，日本研究に焦点をあてたサイト[2]が活動を行っている．また，データベースが整備されると同時に，数理的手法による研究が増え，より客観的視点に基づいた議論が行われるようになってきた [3, 4, 5, 6]．

筆者らも数理的手法のための基礎とするデータベース開発を進めてきたが，研究を進めるうちに，原典や翻訳のデータベースはもとより，歌の修辞技巧データ (たとえば，序詞，掛詞のある歌番号とその種類，分類，係結びの歌番号と係結びの種類，縁語) や注釈のデータや地名，人名のデータ，品詞タグ付データ，数理計算をするために行った前処理データ，分析結果をまとめたデータなど，多変量解析に利用するさまざまなデータが発生した．しかしながら，これらのデータは，個人的に数値処理を行って結果を出し，論文を書くだけでなく，データベースとして管理，公開しておけば，他の領域を同じくする研究者にとっても有用なはずである．

そこで，1) 原典データの加工，既存の資料の電子化【基礎】，2) 原典の情報をさまざまな視点からサポートあるいは詳細化するデータの開発【詳細】，3) 数値処理が行える形式への前処理【前処理】を軸としてデータベース化を進めた．

具体的には，1) 国文学研究資料館開発のデータベース [2] を基礎データとして利用[3]，2) 英翻訳本 [7, 8, 9] を電子化し，翻訳データを作成，3) 品詞タグつきデータ，4) 修辞技巧タグつきデータ，5) 近藤みゆき氏提供のジェンダーデータ[4]を追加し，後述するシステムでWEB公開できるようにした．以下では，そのうち翻訳データ，品詞タグつきデータの開発について述べる．

2.1 翻訳データ

素材：英翻訳データは３種類電子化し，データベース化した．まず，はじめに許諾を得，Rodd[7] の翻訳を電子化し，これを公開用データとすることとした．つぎに，McCullough[8] の翻訳も翻訳者の違いが結果に及ぼす影響を検討するため，同様に電子化し，タグ付を行った (非公開)．最後に，日本人による英語翻訳が上記２翻訳とどのように異なるかを比較するために，Honda[9] についても同様の作業を行った．

構造：表１に示すようにデータは１首１レコードとし，IDを与えた．Rodd[7]，McCullough[8] はできるだけ５７５７７に対応するよう翻訳してあるが，必ずしも日本語の各行とは一致してはいない．Honda[9] は，まったく５７５７７を意識せず翻訳している．３翻訳とも原文通りに入力し，改行のある箇所をスラッシュで示した．ただし，Honda[9] は古語英語を随所に用いており，別フィールドを用意し，現代英語表記に改めたデータも追加した．最後に翻訳中の英単語に品詞タグをつけ，英翻訳データ

[1] 日本研究・日本文学に関する学会が国際会議として海外で定期的に開かれるようになった．たとえば，EAJS (European Association for Japanese Studies) の国際会議は 1976 年以来 3 年に 1 度開催されている．中でも 2000 年フィンランドでの開催には，33 カ国から約 400 名の研究者が参加し，何とそのうち半数以上の約 270 名が外国人研究者であった．日本研究の研究者は日本人とは限らない時代となっていることがわかる．日本研究が論文という形式だけでなく，これら外国人研究者による翻訳データベースという形式も広く流通する可能性もある．

[2] たとえば，Japanese Text Initiative：バージニア大学エレクトロニック・テキスト・センターとピッツバーグ大学東アジア図書館が共同で進めているプロジェクトで，日本古典文学の電子テキストが WWW で利用できるようになっている．

[3] 同研究プロジェクトによる標準表記をはじめ，おおむね，異本，校訂に関する取り扱いは同じである．

[4] 男性歌人，女性歌人，読人不知のタグ付データ

ベースでの検索でも品詞検索が行えるようにした．またこれにより，日本語にはない事物の単数形複数形の異なりを計算することができるようになった．ただし，現時点では各単語は基底形 (動詞なら原型，名詞なら単数形) へ変換されていないので，意味上の語彙の頻度集計は行えない．これは今後の課題としてデータを追加する．

表 1 Rodd の英翻訳のデータ：古今和歌集巻一 春哥上 ードル記号＋アルファベット 1 文字＋縦棒をタグとしている．$A| は通し番号で体系本番号に相当，$B| は作者名，$C| は前文，$D| は和歌，$E| は品詞タグデータ (以下の略語解説はここにあるもののみ掲載)

```
$A|000001
$B|Ariwara no Motokata
$C|Written when the first day of spring came
   within the old year.
$D|spring is here before /
   year's end when New Year's Day has /
   not yet come around /
   what should we call it is it /
   still last year or is it this
$E|spring/NNS is/VBZ here/RB before/IN
   year's/JJ end/NNS when/WH New/JJ
   Year's/JJ Day/NNS has/VBZ not/NEG
   yet/RB come/VB around/IN what/HW
   should/MD we/PRP call/VB it/PRS is/VBZ
   it/PRS still/RB last/JJ year/NNS or/CC
   is/VBZ it/PRS this/DTG
```

NNS=普通名詞単数形，NNP=普通名詞複数形，
VB=動詞原形，VBZ=動詞 3 人称現在単数形，
RB=副詞，IN=前置詞，　JJ=形容詞，WH=wh 語，
NEG=否定語，MD=助動詞，PRS=代名詞単数，
PRP=代名詞複数，CC=等位接続詞，
DTG=決定詞/代名詞

2.2 品詞タグつきデータ

単語分割の意義：古典文学を計算機で分析する研究には，近藤 (n-gram)[5]，竹田 (LCS)[6] のように計算アルゴリズムによって，数値化し，単語分割作業を必要としないものと，宮島ら [10]，村上ら [11]，村田ら [12] のように語彙の計量を目的とし，そのため単語分割作業を前提とするものに大きく分かれる．

前者のような単語分割作業を必要としない方法論は単語分割や品詞づけの作業に労力を要すること，専門家間でも品詞の解釈[5] や文法の取り扱いに基本的な考え方の違いがあること[6]，複合語の認定については揺れがあること[7]，などの問題点を解決するものとして注目される．

一方，後者のように語彙の計量を研究目的とする場合には，単位切り，単位認定は避けられぬ作業となる．また，作家の特徴を抽出しようとしたとき，なるべく文章の内容と関連性の薄い要素を用いた方がよい (金 [15]) ことが報告されている．このような場合には，作品内容に依存した意味的特徴よりも，文法のような内容に依存しない特徴抽出を用いるほうが妥当であり，そのためには単語を品詞ごとに分けて，統計処理を実施したほうがよいと思われる．

長期的視点という要素はあるが，データベースのための単語分割あるいは品詞データ作成作業を行う意義について，次のように考えた．

1. 確かに労力はかかるが，一度データベースを作成すればずっと使える **【苦の解決】**
2. 品質管理を丁寧に行えば，いずれ実用に耐えられるようになる **【質の解決】**
3. 古典作品は今後ともどんどん増え続けるデータとは異なり，作業は有限．共同開発すれば，作業時間も短縮できる．**【量の解決】**

本研究の目的は，和歌に含まれる特徴や文化を抽出し，それが異なる文化を背景とする研究者のそれとの差について数量解析するものである．したがって，筆者らは，宮島ら [10] の分割単位にならい，語彙の計量ができるよう，各和歌を単位分割し，品詞タグをつける作業を開始した．

方法：古語辞書を作成し，その辞書を使って置換を行うプログラムを作成した．これで作業のあらかたを行い，細かな修正は人手によって行った．ただし，プログラムは文脈ルール，いわゆる文法解析は行わず単純に最長一致文字列置換のみの機能とした[8]．

具体的な手順は次のようになった．まず，宮島ら [10] のデータおよび ATOK8/9 用古文入力変換率向上支援単語ファイル[9] を加工して置き換え用古文単

[5] 「はるくれば かりかへる なり(古今 30)」の伝聞推定の「なり」をめぐって，「ラ変型活用語には連体形に，四段動詞には終止形につく [13] (p.24)」とあるが，一方で「かりかへるなり.. 「なり」はすべて連体形につく．終止形につくというのは誤解 [14](p.88)」とも．

[6] 岩波古語辞典では (1) 動詞は終止形見出しではなく，連用形見出しとし，(2) 形容動詞は認めない方針が貫かれている．したがって，(古今 562) の「けに」は岩波では形容動詞ではなく副詞であり，宮島ら [10] では形容動詞を品詞として認めているため「け, 異, 形動」となっている．

[7] 近藤 [5] は「一語をどう認定するかは，その基準の立て方にも様々な立場があり，従来から多くの研究がなされてきた．そもそも単位をめぐる基準からして，一通りではない」と述べ，複合語処理の難しさを説明している．

[8] 一般的に語および活用形の文法を用いると多くの情報を利用することになり，変換効率がよくなるように考えられるが，古典文法はいつも同じではなく時代によって異なる．

[9] Ver. 1.1 太田瑞穂氏作．ワープロ古文入力のための辞書．動詞・助動詞・形容詞の活用形などのエントリを持つ．一般的に，語彙計量では動詞活用形のそれぞれを集計したり，助動詞

語辞書を作成した．岩波体系本表記のデータを作成し，最長文字列一致で，単語[品詞-詳細]形式のデータに置き換えられるプログラムを作成した．学習参考書を利用し，動詞助動詞など頻繁に用いられる述部パターンを登録した[10]．

以上，機械的手続きによる大雑把な作業で，以降は手作業で，品詞タグのついていない箇所のタグづけ作業，係結びタグの付与，読み，活用形の誤りの修正，品詞とは関係ないが，序詞，掛詞のフィールドを追加し，その記述を行った．

その際，辞典および各種文法解説書[13]を利用した．また，諸説の見解をできるだけ考慮するため，片桐[16]，久曾神[14]を参考にした．片桐[16]の現代語訳はその解説にもあるように，できるだけ語を省略せず，各語を忠実に翻訳していることから，単語や品詞の認定，結びの省略など，明示的でない要素の記述に役立った．

さらに，著作権上公開はできないが，品詞詳細タグを付すために参考とした現代語訳として，久曾神[14](全部)および片桐[16](一部)の現代語訳の入力も行った．品詞および活用タグは表2のように略号で記述した．

表 2　品詞のデータ化：現代語訳と一部に序詞および掛詞が作成作業を通して追加された．Aは歌番号，Bは品詞データ，品詞データ中の字句は全てひらがな．括弧内は品詞が略号で示され，ハイフンで活用形，漢字表記，助動詞・助詞の用法，種類など詳細が記述されている．Cは久曾神[14]現代語，Dは片桐[16]現代語，Eは修辞技巧．データ中の改行は任意．

```
$A|000113
$B|はな [名] の [格助]、いろ [名] は [係助] ／
うつり [ラ四-用] に [完-用] けり [詠-終]
な [終助-感] ／いたづらに [形動ナリ-用] ／
わ [代] が [格助] み [名]、よ [名] に [格助]、
ふる [ハ下二-体-経る/ラ四-体-降る] ／
ながめ [名-詠め/名-長雨]、せ [サ変-未]
し [過-体]、ま [名] に [格助] ／
$C|美しい花の色はいつしか色褪せてしまった
ことよ。いたずらに長雨が降り続いていたうち
に（私の容色も衰えてしまったことよ。なすべ
きこともなく、世を過ごす物思いをしていた間に）。
$D|花の色は褪せてしまっていたのだなあ、虚し
くも。長雨に降りこめられ、また、我が身が世に
暮らしてゆく上での物思いに耽っていた間に。
$E|序詞=わがみよにふる-ながめ，
掛詞=ふる（降る）-ふる（経る），
掛詞=ながめ（眺め）-ながめ（長雨）
```

の計量は行わないので，本来はこれら辞書エントリは作成する必要があった．

[10] 例)「てなりけり」→て [接助] なり [断-用] けり [詠-終]

3　マネージメントシステムの開発

3.1　設計仕様

データベースのモデルのなかで，いくつかの情報を統合する方法としては，次のようなものが考えられる．1) 原典にすべての情報をマークアップする方法(SGML,XML)，2) 個々のファイルは小さな単位としてリレーショナルデータとして持つ方法(SQL)，3) 単なるフラットなファイルとして持ち，相互にリンクする方法，

3) の方法は，一番データベースらしくない方法であり，単なる何の構造もマークアップもないテキストの寄せ集めである．しかし，これはフルテキストデータベースなどの呼び名で使われており，実際にサーチエンジンなどのデータベースはこれにあたる．極端ではあるがファイルであれば何でもよいゆえ，そのまま，CSV データを入れて，縦計算をする，HTML を入れて相互にリンクする，pdf をデコードして版下データベースとする，などいろいろ応用しやすい．

筆者らは，最もシンプルな 3) の方法を基礎に，特別なソフトウエアやまたその知識を必要とせず，定義，タグ付が簡単でわかりやすい方法を検討した．また，データベース操作においては，研究者の行動の一部として，アップロード，公開，利用，ダウンロード，更新，アップロードといったサイクルが円滑に容易にできる方法を検討した．そこで，表3 にしめすような仕様とした．

表 3　システムの概要

1. データの作成に特別なソフトウエアを必要としないこと．
2. 公開が簡単であること．
3. 公開と同時に簡単に数値の計算・集計ができること．
4. スクリプトあるいは，データ構造，計算および検索の定義が簡単にできること．
5. 複数のデータベースが登録できること．
6. 同じ定義であれば，複数のデータベースをまたがって検索，計算ができること．いいかえれば，データベース間の通信ができるような仕組みを持つこと．
7. 複数のデータベースが相互に参照できること．
8. データの訂正やデータの記述方法の変更が発生した場合，それを一意に変更できること．
9. 分析の基となったデータがデータベースオーナーだけでなく，ユーザ二よってもダウンロードでき，それを加工して新たなデータが作成できるようにすること．

以上から，システムを 1) データベースの公開を支援するシステム (BBDB) と 2) 作成や品質の管理を支援するシステム (BBQC) の 2 部で構成することにした．以下に詳細を報告する．

3.2 検索・公開システム (BBDB)

概要：データベースの公開・検索を容易に実施できるように設計されたシステムを BBDB(Bare Bone Database) [11] という．

BBDB は，総合研究大学院大学で公開されている貝塚データベース，小松左京コーパスなどの公開システムをもとに，プログラミングレスでデータベース管理者が簡単にブラウザからアップロードするだけで，公開・検索できるようにしたものである[12]．

BBDB によってデータベースを公開するには，定義部とデータ部を含むファイルを記述する．これらは，簡単なテキストファイルなので，ワープロ，エディタで記述できるものである．

機能：管理者用の WEB 公開マネージャ，ユーザ用の検索，集計，ダウンロードなどの機能をがあり，WEB 公開マネージャとして，登録データベース一覧，データベース登録，削除，ディレクトリ認証管理，アピアランス管理，データベース概要ページ生成，検索指定画面生成，ダウンロード指定画面生成，検索として，一般語句検索，前後語句ソートつき KWIC 検索，数値比較検索，絞り込み検索，AND/OR 検索，集計として，頻度集計 (正逆順位ソート付)，基礎統計 (個数，平均，最大，最小，標準偏差)，ダウンロードとして，ファイル形式変換 (Tab, CSV, BBDB)，テーブル略号変換，文字コード変換，ダウンロード必要フィールドの指定などの機能が利用できる．すべての操作はブラウザから行う．

定義部：古今集 DB を例として表 4 に示す．ただし，簡単のため，これは一般公開用に作成された単純データベースのものであり，複数の相互参照の定義は含まれていない．定義部ではデータの構成・属性のほかに検索や計算の指定が記述されている．情報はすべて$$DB_NAME| のように【ドル２つ＋定義文字列＋｜】で示されるタグで，定義される．

データ部：ファイルのデータ部の開始箇所には，$$DATA| を置き，それ以降にレコードを書くことができる．BBDB 形式の１レコードは，表 5 に示す形式である．各フィールドは，【$A|】のように，ドル１つ＋アルファベット１文字＋縦棒の３文字で始まる．デフォルトでは改行は無効で，各行の適当なところで改行をいれてよい．各フィールドのタグはわずか３文字なので，冗長性も少ないゆえに，誤入力も起こりにくい．閉じタグはない．制約は，１レコード最大フィールド数が A-Z の 26 であること[13]，A フィールドは必ず６桁のレコード ID であること，の２点だけである．

公開手順：定義部データ部の記述されたデータファイルが完成したら，次は，公開処理を行う．すべての公開処理はブラウザを使って行われる．簡単にいえば，ブラウザからデータファイルを転送し，その確認をしたら，ボタンを押してデータ保存する．もう一度ボタンを押して，HTML を生成して終りである．

はじめてアップロードする場合は，定義部とデータ部の全部もしくは一部が記載されファイルをアップロードする必要がある．追加アップロードする場合には，定義部は不要であり，データ部相当を記述したファイルとして，Tab 形式，CSV 形式のファイルをアップロードすることができる．

公開手順を実行することによって，データベースの概要や検索，計算，ダウンロードの各種設定が，自動的に行われる．データベースの概要が生成された画面を図１に示す．１レコードは１HTML 変換[14] し，公開すべきすべてのデータの生成を終了する．このようにあらかじめ設定ファイルから表示用のファイルまで，ボタン一つですべて作成してしまう．

表 4　古今集 DB の定義ファイル (一部)

```
$$DB_ID|KW
$$DB_VER|2.0
$$DB_NAME|古今和歌集データベース
$$DB_OWNER|Hilofumi Yamamoto
$$DB_EMAIL|yamagen@ucsd.edu
$$DB_ABST|
古今和歌集 1,111 のデータベースで，歌一つが１レコードで収録
されている。歌の原表記，仮名表記，岩波仮名表記，ローマ字表
記，英語翻訳，品詞解析データ，作者名，作者名標準表記，作者
性別などが収録されており，作者別の集計や性別による歌の分類，
分析ができる。
$$HEADER|
$A|歌番号=歌につけられたユニークな番号 (6 桁)
$B|作者=本文に見られる作者名
$C|作者標準=作者の標準表記
$D|性別=作者の性別
( 検索は m ＝男 / f ＝女 / n ＝読人不知 を指定すること )
$E|作者英文=作者の英文表記
$F|題=各歌の題
$G|題仮名=題の仮名表記
$H|題英語=題の英語表記
$I|歌=歌の標準表記
$J|歌仮名=歌の仮名表記
$K|歌岩波=歌の岩波体系本による仮名表記
$L|歌品詞=歌の品詞分類
$M|歌ローマ=歌のローマ字表記
$N|歌英語=歌の英語翻訳
$O|解釈=英文 (Rodd,L.R.) による解釈
$P|リンク=歌人データベースへのリンク
$$REPLACE|
$D|<img src="../image/KW_%%VAL%%.gif" alt="%%VAL%%">
$$FIELD|
```

[11] Bare Bone は，直訳すれば「骨むき出し」の意だが，筆者らは，飾りなど一切ないが，シンプルであるがゆえに誰にでも使いやすく，かつ重要なもの，を意図している．

[12] 2001 年，及川が筑波大学で行っている数理考古学の授業において BBDB はデータベース作成の演習に利用され，受講生は夏休みの宿題として，各 50 レコード以上の考古学関連データベースを作成した．

[13] 26 以上使えても使えなくても，それよりもまず，１レコードが本当に 26 の要素を持たなければならないものなのかを，検討するべきであろう．

[14] これを我々は 1R1H と呼んでいる．

表 5　古今集 DB の 1 レコード (000007)：一般公開版につき，作者の公開許諾のないデータはここには掲載されていない．各フィールドの詳細は，表 4 の HEADER 定義にしたがっている．

```
$A|000007
$B|よみ人しらす
$C|読人不知
$D|m@
$E|Anonymous
$F|題しらす
$G|たいしらす
$H|Topic uuknown.
$I|心さし／ふかくそめてし／おりけれは／きえあへぬ雪の／花とみゆらん@
$J|こころさし／ふかくそめてし／おりけれは／きえあへぬゆきの／はなとみゆらむ@
$K|こころざし／ふかくそめてし／をりければ／きえあへぬゆきの／はなとみゆらむ@
$L|こころざし-名@／ふかく-形ク-用@，そめ-マ下二-用@て-接助@し-副助-強@／をり-ラ四-用@けれ-詠-已@ば-接助@／きえ-ヤ下二-用-消える@，あへ-ハ下二-未@ぬ-消-体@，ゆき-名-雪@の-格助@／はな-名@と-格助@，みゆ-マ上一-終@らむ-現推-体@／@
$M|kokorozashi / fukaku someteshi / orikereba / kieaenu yuki no / hana to miyuran /
$N|so longingly have I / awaited the fresh flowers / of spring that they have
 / dyed my soul  and I see snow / as clustered blooms on branches /
$O|The former Chancellor was Fujiwara no Yoshifusa
```

3.3　品質管理システム (BBQC)

概要：公開前の入力ミスなどチェック，公開後の更新，追加などこれまで多くのユーザが手作業で行っていたさまざまな品質管理の省力化，自動化を目的とし，実施できるように設計されたシステムを BBQC(Bare Bone Quality Control) という．標準的な機能の一覧は表 6 に示す通りである．BBQC による品質管理処理は，ユーザが BBDB 形式のファイルを転送し，サーバ上で処理した後，ダウンロードもしくは BBDB に転送し，完了する．

機能：ファイルをサーバに転送するとまず，定義部の文法チェックが行われる．その後，定義部にしたがって品質管理のための処理が行われる．表 6 の C1-10 は品質管理のためのチェックポイントである．文字種のチェックや不要な空白，数値のタイプ，範囲などが機械的にチェックされる．また，重複 ID，欠番 ID のチェックの機能，チェックにより判明した不都合をサーバ上で修正する機能 (一括項目名変更，手作業修正)，また，その修正を確認するためのファイル内容表示，ファイルの追加，結合，不要になったファイルの削除機能，ユーザが自分のパソコンの表計算ソフトで作成した，csv, tab 形式のファイルを bbdb 形式に変換する機能，文字コード変換，データ項目一覧作成機能，などを持っている．

システム内部では，それぞれの仕事に応じたフィルタプログラムが定義部にしたがって実行する．仕組みとしては簡単であるが，実際，これらすべてのチェックポイントについて手作業で行うのは簡単ではない．

表 6　データベース品質管理操作メニュー

```
A. upload definition file
B. upload data file
C. preset check points
      1. 1-byte char
      2. 2-byte char
      3. unnecessary space deletion
      4. integer
      5. float
      6. unknown table code
      7. mandatory
      8. field duplication
      9. max. number
     10. min. number
D. process data
   a) quality control
      1. quality check (checking point)
      2. duplicating/missing id check
      3. item authorization check
   b) data modification
      4. replace values
      5. modify record by field item
   c) file handlings
      6. delete file
      7. cat file
      8. append file to file
   d) file code and format conversion
      9. tab -> bbdb
     10. bbdb -> tab
     11. csv -> tab
     12. tab -> csv
   e) item list generation
     13. in alphabetical order
     14. frequency in use of a word
E. modify record on the web
F. delete record on the web
```

4 データベースの利用の実際

実際には，原典となる古今集 DB，3種の翻訳データ，品詞データ，作者データの4点からなるデータベースが作成されているが，著作権上公開できないものがあるので，実際に公開しているサイトを例に，BBDB と古今集 DB の利用の実際について説明する．

古今集 DB のページ[15] に行くと，図1に示すように定義部に記述した内容が表示され，データベースの概要を知ることができる．上の検索リンクを選んで，検索を行う．図2にあるように，語句検索 (and, or)，KWIC 検索，数値比較検索を混在させた絞り込み検索ができる．

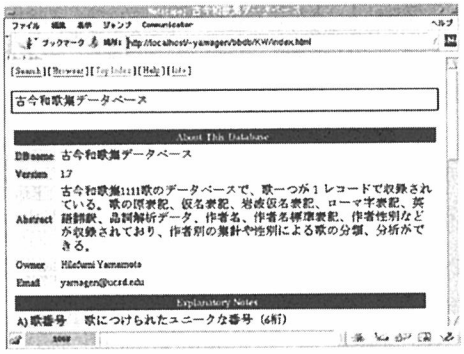

図 1　生成された概要ページ

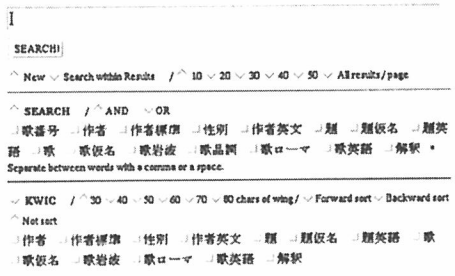

図 2　古今集 DB の検索指定画面

たとえば，和歌の中で「鳥」を詠んでいる歌を検索したいとする．すると結果表示画面3では，30件ヒットしたことがわかる．この画面には，それぞれのレコード (HTML) へのリンクと次の10件を表示するためのリンクとダウンロード用のリンク [download]，統計計算のためのリンク [stat] が用意されている．リンク KW000143 をクリックすると，素性の歌のページへ行く．もちろん，これはあらかじめ生成された静的なページであるので，BBDB 以外からもリンクすることができる．ここで，[download] をクリックすると，「鳥」30件のデータがダウンロードできる．その隣の [stat] をクリックすると，大雑把ではあるが頻度と基礎統計の計算ができる．たとえば，作者標準の FREQ を ON にして，[stat] ボタンを押す (図4) と，表7のように27の歌の作者分布がわかる[16]．さらに歌品詞を指定し，集計した結果が表8である．これによると，鳥30羽のうち12羽が「ほととぎす」であることがわかる．

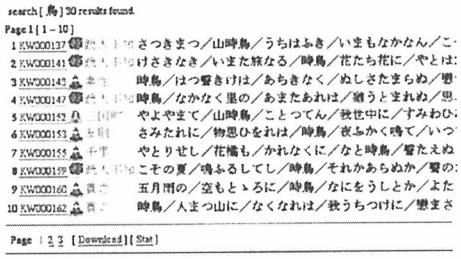

図 3　「鳥」の検索結果

図 4　基礎統計データ出力指定画面

表 7　基礎統計データ出力結果「作者標準」

0	27	100.00%	TOTAL
1	13	48.15%	読人不知
2	3	11.11%	友則
3	2	7.41%	貫之
4	2	7.41%	忠岑
5	1	3.70%	業平
6	1	3.70%	三国町
7	1	3.70%	敏行
8	1	3.70%	寵
9	1	3.70%	千里
10	1	3.70%	躬恒
11	1	3.70%	素性

[15] http://aci.soken.ac.jp/ kokinshu/KW/

[16] 結果出力で，"30 found" と出ていたのは，30箇所に「鳥」が見つかったの意味．

表 8　統計データ出力結果「歌品詞」(一部)

0	397	100.00% TOTAL
1	21	5.29% の-格助
2	15	3.78% に-格助
3	12	3.02% ほととぎす-名
4	10	2.52% も-係助
5	8	2.02% を-格助
6	8	2.02% ば-接助
7	8	2.02% と-格助
8	7	1.76% が-格助
9	6	1.51% て-接助
10	5	1.26% は-係助

5　おわりに

本論[17] では，古今集 DB の品詞データと英翻訳データの開発について報告した．品詞データには，まだ誤りや活用形の未整理の箇所があり，今後もデータを見直し，修正を加えていく必要があるが，品詞や活用形の認定，複合語の取り扱いが研究者によって異なり一意に決められず，データ化する上で問題となることを述べた．

また，公開や品質管理を効率的に行うシステムを開発し，その目的，利用方法，利用例について報告し，定義ファイルとデータファイルをアップロードするだけで，検索，集計処理の指定まで一括して行え，簡単な計算処理まで行えることを示した．

今後の課題として２つのことがある．ひとつは古今集 DB のコンテンツの充実とその精度の向上，もうひとつは BBDB，BBQC の公開と普及である。

前者は，現在の DB を広く公開することによって，利用者である研究者からデータベース中の誤りを指摘してもらったり，新たな意見や提案をデータベースに反映することによって実現していくことが可能である．そのためのツール群も BBDB の中に組み込んでいく予定である．

後者については，すでに公開のためのプラットフォームとなるサーバを総合研究大学院大学の図書館に設置しており，マニュアルの作成と並行してその準備作業を進めている．BBDB，BBQC を広く普及していくためには，まず，研究者が手軽に利用できるプラットフォームが必要で，今後研究者への呼びかけやワークショップの開催などを通じて，その実現を図っていきたいと考えている．

[17] 国文学研究資料館の中村康夫先生には国文学研究資料館データベース二十一代集の利用を快諾していただいたばかりでなく，CDROM を貸与してくださいました．国文学研究資料館の安永尚志先生には，パラレルテキストと国文学資料について御指導いただきました．青山学院大学の近藤泰弘先生には，品詞データ作成の際，国文法についてご指導いただきました．実践女子大学の近藤みゆき先生には，ジェンダーデータつき単位切りデータを提供していただきました，コロラド大学ボールダー校のローレルラスプリカロッド先生にはご著書をいただいたばかりでなく，データベース化および公開について，快諾いただきました．お礼申し上げます．

参考文献

[1] 佐竹昭廣, 立川美彦: 重層型情報時代に対応する国文学高機能情報形成手法の開発とその実用化に関する研究, 技術報告, 国文学研究資料館 (1998). 平成７年度～平成９年度科学研究費基盤研究 (A)(2) 研究成果報告書 (課題番号 07401014).

[2] 中村康夫, 立川美彦, 杉田まゆ子: 国文学研究資料館データベース 古典コレクション『二十一代集』(正保版本) CD-ROM, 岩波書店 (1999).

[3] 村上征勝: 文章分析と統計学, 数理科学 特集 知としての統計学, Vol. 11 月号, No. 389, pp. 27–33 (1995).

[4] 近藤みゆき: n グラム統計処理を用いた文字列分析による日本古典文学の研究—『古今和歌集』の「ことば」の型と性差—, 千葉大学「人文研究」, Vol. 29, pp. 187–238 (2000).

[5] 近藤みゆき: n-gram 統計による語形の抽出と複合語—平安時代語の分析から—, 日本語学, Vol. 20, pp. 79–89 (2001).

[6] 竹田正幸, 福田智子, 南里一郎, 山崎真由美, 玉利公一: 和歌データからの類似歌発見, 統計数理, Vol. 48, No. 2, pp. 289–310 (2000).

[7] Rodd, L. R. and Henkenius, M. C.: *Kokinshu - A Collection of Poems Ancient and Modern*, Cheng and Tsui Company, Boston MA USA (1984).

[8] McCullough, H. C.: *Kokin Wakashu, The first Imperial Anthology of Japanese Poetry Translated and annotated by Helen Craig McCullough with Tosa Nikki and Shinsen Waka*, Stanford University Press, Stanford, CA, USA (1985).

[9] Honda, H.-H.: *The Kokin Waka-Shu; The 10th Century Anthology: edited by The Imperial Edict*, Hokuseido Press, Tokyo (1970). Translated by H. H.Honda.

[10] 宮嶋達夫, 中野洋, 鈴木泰, 石井久雄: フロッピー版古典対照語い表, 笠間書院 (1989).

[11] 村上征勝, 今西祐一郎: 源氏物語の助動詞の計量分析, 情報処理学会論文誌, Vol. 40, No. 3, pp. 774–782 (1999).

[12] 村田菜穂子, 岩田俊彦: 平安時代の文学作品における形容動詞対照語彙データベースの構築とそれを用いた語彙論的研究, 情報処理学会研究報告, Vol. CH45–10, pp. 73–80 (2000).

[13] 小林和彦: 古典新釈シリーズ 5 古今和歌集, 中道館 (1978).

[14] 久曾神昇: 古今和歌集 全訳注 (1)-(5), 講談社学術文庫 (1982).

[15] 金明哲: 自然言語における統計手法を用いた情報処理, 統計数理, Vol. 48, No. 2, pp. 271–287 (2000).

[16] 片桐洋一: 古今和歌集全評釈 上・中・下, 日栄社編集書 全３巻, 講談社 (1998).