人文科学とコンピュータ 54－5
（２００２．５．３１）

古今和歌集データベースの開発と和歌の数理解析*

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 DB) 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 の和歌データを含むほか，英語翻訳，品詞解析データも作成し，これに追加した．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) システムから利用できるデータ分析，について述べる．

Development of Kokin Waka Shū Database for Mathematical Analysis

Hilofumi Yamamoto† Akifumi Oikawa‡

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

Abstract

We have developed the Kokin waka shu database (Kokinshū DB), 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) Possibilities of mathematical analyses using data extracted from the 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

古今和歌集は，10 世紀末に成立した，勅撰和歌集である．この歴史ある作品の研究者人口は，一体どれぐらいなのだろうか．和歌の研究者だけでなく，古典文学の研究者，和歌の愛好家，さらに海外の日本古典文学研究者を含めると，その数は決して少なくなく，また，その研究分野も広いはずである．

データベースをデザインする上で，どのような研究領域があるのか,「古今集」をキーワードに国語学研究文献総索引データ [1] で検索・集計したところ，表 1 のように文章・文体，文法に関する研究が多いことがわかった．作品の鑑賞が中心と考えられそうだが，この作品には実にさまざまな研究領域が存在することがわかった．

表 1:「古今集」の研究分野

頻度	領域	頻度	領域
52	文章・文体	7	音声・音韻
21	古典の注釈	5	年鑑単行本
20	国語史	4	国語資料
18	大辞典	2	文字・表記
18	語彙・用語	2	国語教育
16	白紙台紙	2	国語学一般
14	文法	1	書評・紹介
11	追補	1	敬語・丁寧語

本論では，次の２点について述べる．

１つは，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文学教育とりわけ，和歌の教育に役立つコンテンツを提供すること．具体的には，Rodd[2] の和歌翻訳を利用し，古今和歌集の 1,111 首の日英パラレルテキストデータベースを開発することである．

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ことから，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．パラレルテキストデータベースは，外国人研究者の古典入門だけでなく，外国人の日本観といった文化研究にも役立つはずである．

もう１つは，外国人研究者を含めて，このようなさまざまな領域のさまざまな研究者が，自ら作成したデータを公開し，検索，集計，あるいは修正，更新に至る処理を容易に実施する仕組みを提供することである．

旧来のカードにかわり，データベースシステムが研究に用いられるようになり，その中で，定義，入力，保存，修正，追加という作業が行われるが，これにインターネット公開となると，アップロード，修正，追加，集計，ダウンロードというサイクルが加わる．実際にこれを実施しようとすると，多くの知識を必要とし，多くの困難に直面する．筆者らは研究者がこのようなサイクルを簡単に日常的に実施できるような，システムが必要であると考え [3]，その開発を行ってきた．

以下では，これらシステムの開発，それを利用した古今集データベースの開発，その利用例として，データベースを使った計算処理，について述べる．

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 [4, 5]，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[1]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 [6, 7, 8, 9]．

本研究では，このようなデータベース時代の恩恵と流れを受け，国文学研究資料館開発のデータベース [5] を基礎データとして利用[2]し，英翻訳データの開発，品詞タグつきデータの開発を行った．開発されたデータは，近藤みゆき氏提供のジェンダーデータとともに，後述する BBDB で公開できるよう，整備した．以下では，そのうち品詞タグつきデータ，翻訳データの開発について述べる．

2.1　品詞タグつきデータ

古典文学を計算機で分析する研究には，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と，近藤 (n-gram)[8]，竹田 (LCS)[9] 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に大きく分かれる．

確かに単語分割の作業は大変な労力を要する上に，専門家間でも品詞の解釈に大きな違いがあるため，非常に困難である．たとえば，

(古今 30：凡河内躬恒)

はるくれば　かりかへる なり

[1]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．

[2]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．

しらくもの　みちゆきぶりに　ことやつてまし

における伝聞推定の「なり」について，小林 [14] は，『雁帰るなり..(略)「なり」は，詞書によって,「雁の声を聞き」，そうと推定した意とわかる．四段動詞には終止形(ラ変型活用語には連体形) に付く伝聞推定の助動詞 (断定「なり」は，体言・連体形に付く)．(略) (p.24)』と述べ，一方，久曾神 [15] は,『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．(p.88)』と述べている．岩波古語辞典では，終止 (ラ変連体) としている．

文法の取り扱いにも基本的な考え方の違いが存在する．岩波古語辞典 (大野晋他監修) 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．(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．

また，複合語の認定の問題，たとえば (古今 563) のような「さえまさり [ラ四-用]」を 1 語とするか，２語とするか，も語彙の計量研究に大きな問題がある．近藤 [8] 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．

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，作者の特徴を抽出する方法として金 [16] は「なるべく文章の内容と関連性の薄い要素を用いるべきである」とし,「単語を品詞ごとに分けて処理したほうがよいという提案がある」と，単位分割の利点を述べている．

筆者らは，これら議論を考慮した上で，宮島ら [10] にならい，語彙の計量ができるよう，各和歌を単位分割し，品詞タグをつけた．この品詞データの作成については，辞典および各種文法解説書 [14] 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 [17]，久曾神 [15] を参考にした[3]．

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

```
KW000001|K| とし [名-年] の [格助]，うち [名-内]
に [格助] ／はる [名] は [係助]，き [カ変-
用] に [完-用] けり [詠-終] ／ひととせ [名-一年]
を [格助] ／こぞ [名-去年] と [格助] や [係助
-疑-係]，いは [ハ四-未-言ふ] む [意-体-結] ／こ
とし [名-今年] と [格助] や [係助-疑-係]，
いは [ハ四-未-言ふ] む [意-体-結] ／|□
```

[3]著作権の関係上公開はできないが，分析および参照のため，久曾神 [15] の現代語訳のデータ化も行った．

2.2　英訳データ

英翻訳データは，許諾を得て，Rodd[2] の翻訳文を公開用データとして開発した[4]．表 3 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[2] はできるだけ５７５５５に対応するよう翻訳してあるが，必ずしも一致してはいない．本記述ではスラッシュで区切ってある．

表 3: 英翻訳のデータ化

```
$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
```

翻訳中の英単語に品詞タグをつけ，英翻訳による検索でも品詞検索が行えるようにした．このタグづけ作業には，Brill's Taggar[19] を用いた。ただし，現時点では各単語は基底形へ変換されていないので，英単語の頻度集計は行えない．

3　マネージメントシステムの開発

一般的に，DBMS というと航空券予約や銀行取り引きのような頻繁に SQL でコントロールするトランズアクションに頑健なシステムを重い浮かべる．しかし，筆者らのシステムはこのようなシステムとは大きくことなる．

人文科学の分野のみならず，さまざまな領域で，データベース公開・共有の利点に関する議論がよく行われる．データベースを公開し，共有することによってよりよい成果を上げていくものと思われる．しかし，具体的な作業として，ブラウザや ftp など公開に必要なツールが便利になったものの，計算機科学を専門としない研究者がこれら作業を実施するのは必ずしも簡単であるとはいえない．また，公開はしたが，更新されず，更新となると再びデータの見直しから始めなければならず，実に修正されないデータがいつまでもサーバに放置されることになる．我々のシステムはアクセス仕様，汎用言語，排他処理，検索スピードといった内部処理やベンチマークといった性能のよさを追求したものではなく，研究者の行動の一部として，アップロード，公開，利用，

[4]また，McCullough[18] の翻訳文も非公開ではあるが，翻訳の比較研究および翻訳の違いがデータ処理に及ぼす影響を検討するため，同様の方法で開発した．

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行えるよう支援するというものである．

以下に，データベースの公開を支援するシステム，BBDB と品質管理を支援するシステム，BBQC の詳細を報告する．

3.1 検索・公開システム (BBDB)

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) [5] 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[6]．

BBDB によってデータベースを公開するには，定義ファイルとデータ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．

まず，定義ファイルについて説明する．古今集 DB を例として表 4 に示す．ここではデータの構成・属性のほかに検索や計算の指定が記述されている．ファイル中の情報はすべて$$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．

次に，データファイルについて述べる[7]．

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＋｜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．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[8]，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．

BBDB の内部動作を図 1 に示す．

すべての公開処理は，ブラウザを使って行われる．はじめてデータベースをサーバにアップロードする場合は，定義ファイルとデータの全部もしくは一部

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|
```

が記載されたデータファイルをアップロードする必要がある．

まず，転送ファイル (trans.bbdb) がサーバに転送される (1) と，定義ファイル (bbdb.def) が出力される (2)．bbdb.def は，データベースの ID，作者，連絡先，概要，バージョン，１レコードに含まれる情報，テーブルコードなど，データベースの内容に関わる情報と，検索方法，計算方法に関わる情報が記述されている．これらの情報はデータベースのトップページとして生成される．表示された画面を図 2 に示す．同時に 1R1H 生成に利用される template.html が出力される．

つぎに，転送ファイル (trans.bbdb) はメインファイル (main.bbdb) としてコピー (3) された後．検索用のテーブル (5:search tables) と検索結果用のインデクス (4:result index) を生成する．

最後に，trans.bbdb から 1R1H(6) を生成して，公開すべきすべてのデータの生成を終了する．この 1R1H というのは，1 Record が 1 Html として，出力されたファイルの意味で BBDB では，あらかじめ表示用のファイルはアップロードした際に作成してしまう．

この間，(3) の直前で，template.html を編集することにより，各レコードの表示の順番やデータベースに記載されていないリンク，組織名の追加などのユーザの好きなページレイアウトを作ることができる．ただし，レコードを追加する時の転送ファイルは，コピーではなく，main.bbdb に追加される．その後，main.bbdb から，検索用テーブルと結果イン

[5] 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．

[6] 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た．

[7] 実際には，定義ファイルもデータファイルも１ファイルとしており，その場合，定義部，データ部とよび，データの開始箇所には，$$DATA| を置く決まりになっている．

[8] 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．

表 5: 古今集 DB の 1 レコード (000007)

```
$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
```

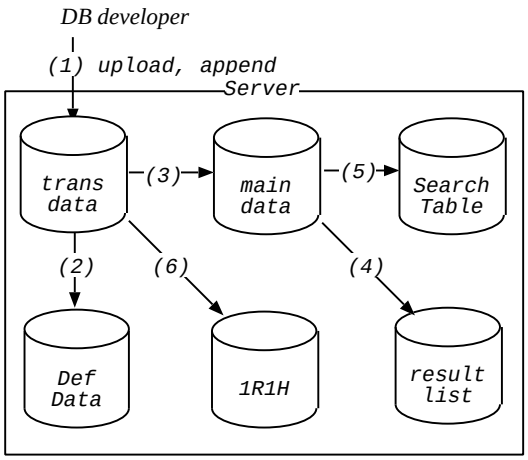

図 1: BBDB による公開手順

デクスが作られ，trans.bbdb から，1R1H が生成，追加される．また，レコードを削除する時の転送ファイル (中身はレコード ID のみ) は，main.bbdb の該当するレコード ID と 1R1H ファイルを削除するのに用いられる．

公開システムには，この他，ユーティリティとして，ディレクトリ毎の認証管理システム，認証パスワード生成システム，ユーザ管理システム，ユーザへの一斉メールシステムなど，公開に必要なものが用意されている．

3.2 品質管理システム (BBQC)

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．

BBQC による品質管理処理は，サーバにファイルを転送し，サーバ上で処理した後，ダウンロードもしくは後述する BBDB に転送し，完了する．

まず，ユーザは BBDB 形式のデータファイルと定義ファイルを転送する．次に，ブラウザからサーバに転送されたファイルの処理を行う．ブラウザには表 6 に示されるメニューがあり，このメニューにしたがって作業を行う．

データ処理は処理のためのオプションの指定 (C) と処理の実施 (D) の 2 つにわかれる．実際のデータ処理では，品質管理に直接関わる a),b) とファイル操作 c),d) が行われる．

(A) で，定義ファイルの文法チェックを行う．定義ファイル bbdb.def はデータの属性が記述されているので，まず，最初に転送されなければならない．(B) で，csv, tab, bbdb の各形式で記述されたデータファイルを転送する．csv, tab の場合は，サーバ転送後，bbdb 形式に変更される．(C) で，C1-10 の品質管理のためのチェックポイントについて設定する．default 値はすでに転送された，bbdb.def から読み取る．ここで指定した値は，ここで再び指定を変更しない限り，D-a) の quality control メニューで利用されつづける．

(C1) は 1 バイト文字以外，(C2) は 2 バイト文字以外，(C3) は不要な空白文字が存在する (たとえば，データ冒頭／末尾など) と，それぞれメッセージを出力．(C4) は数値フィールドで整数以外，(C5) は数値フィールドで小数以外，(C6) は TABLE で指定されたコード以外が存在するとメッセージ出力する．(C7) は MANDATORY 指定されたフィールドの有無のチェック，(C8) はコンマ区切りのフィールドで，

同じフィールドに同じ値が存在するとメッセージを出力する．内部では，ソート，ユニーク処理を実施し，修正を行う．(C9) は数値フィールドで最大値以上が，(C10) は数値フィールドで最小値以下が，それぞれ存在するとメッセージを出力する

D1-3 では品質管理のための調査コマンドを実行する．(D1) は上記 C1-10 に該当する箇所をストリームチェックする．(D2) は重複する ID，欠番となっている ID があるかどうか，チェックする．(D3) は辞書，リストなどある基準を持つ外部ファイルを参照し，その中に存在する項目名であるかどうかをチェックする．外部ファイルはあらかじめ転送しておかなければならない．

D4-6 は，D1-3 にて判明した不都合をサーバ上で修正する．(D4) は一括して項目名を変更する．(D5) はフィールドを指定し，ブラウザ上，手作業で修正する．

D6-8 は，上記，a),b) を実施するのに必要な一時ファイルの作成，内容の表示，追加，結合などの加工や不要となったファイルの削除をおこなう．

D9-12 は，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 形式に変換するためのものである．

D13-14 は，D3 のオーソリティファイルとして生成するもので，項目を (D13) はアスキー順で，(D14) は頻度順で，ソートする．それぞれ「頻度データを添える／添えない」,「頻度 n 以上／以下を出力」のオプションがある．(E) は，1 レコード単位での変更を WEB 上で行い，新規レコードとしてレコードの追加も行う．(F) は，不要となったレコードを削除する．

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ファイルにしたがって実行する．仕組みとしては簡単であるが，実際作業として，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．

4　システムの利用とデータ分析

BBDB はユーザに対して (1) データベースの概要，フィールド情報の提供，(2) 検索，(3) ダウンロード，(4) 集計，のサービスを提供する．

(1) は，図 2 に示すように定義ファイルに記述した内容が整形して出力されたものである．検索は，図 3 にあるように，語句検索 (and, or)，KWIC 検索 (両翼長さ指定，前後の語句ソート)，数値比較検索，の 3 種類をそれぞれ混在させた絞り込み検索ができる．ダウンロードは，全データと検索絞り込みデータの 2 種が利用できる．集計は，データの項目

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
```

毎の頻度および基礎統計 (平均，S.D.，最大，最小) が利用できる．それぞれのサービスのどれをユーザに提供するかは，DB 管理者が定義ファイルで指定する．

検索からデータ分析までの流れを眺めると図 4 のようになる．

ユーザは，検索画面より検索を行う (1) とシステムは，db table をサーチし，該当レコードをあらかじめ作成していた，Result Index ファイルから出力する (2)．Result Index 表示画面 5 には,「鳥」を検索し，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 (1R1H) 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 (1R1H) へジャンプする (3)．この段階で，[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．このデータにしたがって分析すると決意している場合は，そのままダウンロードしたファイルを加工して，任意の統計

図 2: 生成された概要ページ

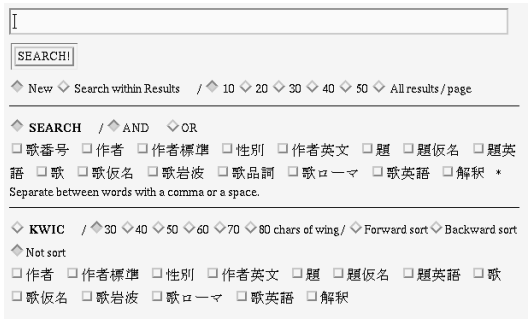

図 3: 古今集 DB の検索指定画面

処理を行えばよい．しかし，期待はずれの場合もなくはないので，隣の [stat] をクリックして，大雑把ではあるが数量の把握を行う．

まず，作者標準の FREQ を ON にして，[stat] ボタンを押す (図 6) と，表 7 のように 27 の歌の作者分布がわかる[9]．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 6: 基礎統計データ出力指定画面

[9] 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．

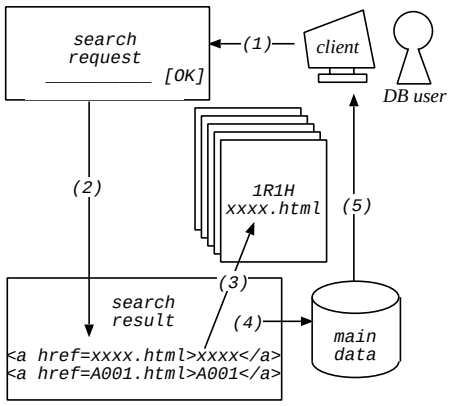

図 4: 検索利用からデータ分析までの流れ

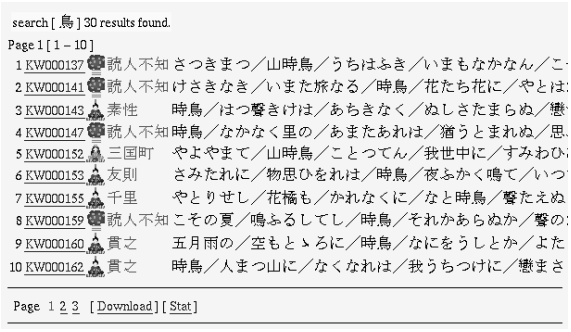

図 5: 「鳥」の検索結果

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

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[10]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．

参考文献

[1] 熊谷康雄他: 国語学研究文献総索引データ, インターネット一般公開版 1.02 (1999-04-23) (1999).

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

[3] 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, 情報処理学会人文科学とコンピュータ研究会, Vol. 49, No. 7, pp. 49–56 (2001).

[4] 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).

[5] 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).

[6] 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).

[7] 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).

[8] 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).

[9] 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).

[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] 竹田正幸, 福田智子, 南里一郎: 歌集間における表現特徴の自動抽出—部分文字列の生起頻度にみる—, 情報処理学会研究報告人文科学とコンピュータ, Vol. CH47–6, pp. 39–46 (2000).

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

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

[16] 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).

[17] 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).

[18] 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).

[19] Brill, E.: Some Advances in Transformation-Based Part of Speech Tagging, Technical report, MIT (1994).

[10]国文学研究資料館の中村康生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．