Vol.2013-IFAT-110 No.3
Vol.2013-DD-89 No.3
2013/2/28

不定型な関連性の格納が可能な和歌データベースの実装

白井 涼子[†1]　波多野 賢治[†2]

概要：近年，和歌研究において，出版社から市販されている和歌データ集やウェブ上で一般公開されている和歌の電子データが利用いられるようになっている．これは，研究対象となる和歌の数が多く，これまで研究者の目に頼っていた手法から計算機を使用して発見科学的に知識を見いだすためである．しかしながら，これらのデータは単に提供されているだけであり，和歌研究で必要となる知見や研究対象となるデータの追記は実質上難しく，和歌研究を進める上では大きな障壁となる．そこで，和歌研究者が必要とする知見やデータの追記などが容易に行うことが可能な和歌データベースの実装を行った．

1. はじめに

和歌集は，古いもので奈良時代から存在し，刷版技術のない時代には和歌集を手で書き写すことにより，後世へ写本という形で伝えられている．このとき，書写者の書き誤りや，書写者があえて表現を変更することがあり，漢字かな表記や，単語の表現が写本によって異なる場合が起こりうる．

和歌研究者の多くは，その異なり部分に何らかの時代的特徴があると考え，和歌が成立した時代背景や言葉の流行り廃りなどを知見として，和歌研究を行っている．この種の研究では，ある和歌集に対する複数の写本に記載されている和歌の表記から，あるルールや知見を発見することが和歌研究を行う上での第一歩とされている．そのため，近年では計算機を用いて発見科学的にこうした知識を見出し，その知識を活用して和歌研究が行われている [9].

このとき一般的には，市販されている和歌データ集 [6][7] や一般公開されている和歌データ[*1]を利用することになるが，データとして提供されている以上のものは当然のことながら扱えず，またデータの追加を行おうにもその方法が和歌研究者にとっては難しい．この困難さは計算機の扱いに長けていないというだけではなく，単に和歌データの追加に留まらず，和歌同士や和歌と各データの関連性をデータ構造の制約が多いデータベースには格納出来ないという困難さも併せ持つ．そのため，和歌からルールや知見を取り出そうにも利用可能なデータや得られる知見は一部に限られ，計算機を用いた和歌研究分野の発展に限界が生じるという問題を引き起こす．

そこで本稿では，以前著者らが提案した和歌データのためのプロパティ jpoem を利用した和歌データベース [10] において，和歌研究者らがデータベースに対し容易にデータや知見を追加できるような仕組みを実装する．

2. 関連研究

本節では和歌データの格納に関する関連研究について述べる．

RDB (Relational DataBase) を用いて万葉集データベースの構成法について提案されている研究がある [8]．RDB はあるデータを複数項目の集合としてリレーションと呼ばれる表形式で表現するデータ管理手法である．キーと呼ばれるレコードを一意に定める値を用いる事で，リレーションごとにまとめられた複数にまたがるテーブルを管理することが可能となっている．この研究では，和歌の写本ごとに表記やよみの異なりである異訓や異同を想定し，さらに書入である左注を含めたデータベースの作成を RDB を用いて行っている．万葉集データの格納は次の四つの方針に基づいている．

(1) すべての写本データを同一形式で格納する．
(2) 和歌本文，題詞 (タイトル)，左注 (注釈・解説) に分けて格納する．
(3) 句ごとに分解し，検索は句単位で行う．
(4) 全体注と部分注に分けてデータベースに格納する．

この研究のリレーションは 15 個に分かれているが，これはこの和歌データベースで扱える関係性は 15 種類であ

[†1] 現在，同志社大学大学院文化情報学研究科
Presently with Graduate School of Culture and Information Science, Doshisha University
[†2] 現在，同志社大学文化情報学部
Presently with Faculty of Culture and Information Science, Doshisha University

[*1] 国際日本文化研究センター | 和歌データベース, http://www.nichibun.ac.jp/graphicversion/dbase/waka.html

ることを意味している．つまり，それ以上の関係性を扱おうとすると， RDB のスキーマの変更が必要となり，和歌研究者が容易に扱う事はできない．また， 15 種のリレーションによる　から語句　を含むような複雑な問い合わせに対しては実用的な検索時間では検索結果を返せず，大量の和歌データを扱うのに適していない．

XML (Extensible Markup Language)[3] を用いて和歌データの構造化を試みている研究も存在する [11]．XML を用いる事で，テキストデータを木構造データとして扱う事が可能となるため，各データの関連性を木構造で表現できればデータの扱いが容易になる．この研究では特に，書入を格納することを重点に置き，和歌データの XML スキーマの提案および，検索システムの実装を行っている．XML は構造化文書で有リレーションよりも複雑な構造をしているが， RDB のようにリレーションが複数に分かれていない分，ある程度の検索時間で和歌検索が可能であった．しかしながら，文献 [8] 同様 [11] においても，和歌研究者による新しい和歌データの追加を考慮しておらず，新しい関連性追加が容易に行えるようなデータ構造ではない．

以上のことから，今までの研究では和歌研究者が容易にデータベースへのデータおよび関連性の追加が可能なデータベースの提案は成されていない．

3. 和歌データの表現

本稿で利用している和歌データベースは，2 節で挙げた問題点と和歌で扱うべきデータが多岐に渡ることを考慮して，RDF (Resource Description Framework)[2] で記述されたデータ (RDF データ) を元に構築されている．

3.1 RDF と SPARQL

RDF とは，ウェブ上に存在するリソースを記述するための枠組みである．類似した機能を持つものとして XML があるが，XML はデータ構造のみを表現可能である一方，RDF はデータ構造だけでなくリソースを用いてデータ同士の関係性を表現することが可能である点が特徴である．

RDF では，主語 (Subject)，述語 (Predicate)，目的語 (Object) の三つの要素 (トリプル) から構成される意味モデルを持つ．このとき，主語はリソース，述語はプロパティ，目的語はプロパティの値をとり，これらの関係は有向グラフで表すことができる．プロパティは関係性を示すもので，語彙と呼ばれる目的に沿ったプロパティの集合が作成されている．

データ参照を行う際，RDF では URI (Uniform Resource Identifier) 参照を行う．この URI は一定の書式によってリソースを示す識別子のことであり，リソースの場所と名前によって表現され，一意に特定することが可能である．

図 1 にトリプルの有向グラフの例を示す．この例では，リソースが「同志社大学」，プロパティが「ウェブサイト」，プロパティの値が「`http://www.doshisha.ac.jp/`」となる．これは，同志社大学のウェブサイトは `http://www.doshisha.ac.jp/` である，という意味を示している．つまり，リソースは説明を受けるもの，プロパティはリソースから見たプロパティの値の意味づけ，プロパティの値は実際にどういったものであるかを示している．

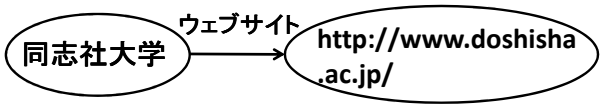

図 1 トリプルの例
Fig. 1 an example for RDF triple

プロパティの値にはリソースだけでなく，文字列をおくことも可能である．述語であるプロパティの値は同時に主語であるリソースにもなることが可能であるため，さらにプロパティが発生してグラフが派生する可能性がある．このとき，一つの語彙ですべてを網羅する必要はなく，複数の語彙を組み合わせることも可能である．このトリプルを組み合わせ有向グラフを拡張していくことにより， RDF データを拡張する事が可能である．

ここで，既存の語彙の例として，代表的な Dublin Core [1] について述べる．

Dublin Core とはウェブや文書の書誌的な情報を記述するための語彙であり，基本となる 15 の要素を用いて意味を表現している．例えば，リソースに与えられた名前を示す `title` やリソースの内容の説明を示す `description` など書式を表現する要素が含まれている．それぞれの要素が広い概念をカバーしているため，さらに詳細に示す場合には定義域や値域も設定され，`title` の正式タイトルの代替である `alternative`，`discription` の目次を指定する `tableOfContents`，そして要約を指定する `abstract` など，細かい指定が可能な拡張プロパティも併せて使用していく必要がある．

この和歌 RDF データに対して検索を行えるクエリ言語の一つが SPARQL [4] である．グラフのパタンマッチングにより検索を行っている．RDB の問合せ言語である SQL に記述形式が類似しており，クエリパタンとして論理積，論理和などを指定可能である．RDF データに対して検索を行うことができることから，ウェブ全体に分散している複数のデータソースに対してクエリの実行が可能である．

さらに，SPARQL の拡張である SPARQL-Update[5] がある．SPARQL ではグラフに対して INSERT や DELETE 機能が付与されていないため，INSERT や DELETE が可能な SPARQL-Update を用いて RDF グラフの編集を行うことが可能である．

3.2 写本内の和歌

和歌研究者が研究に使用する和歌データは，基本的に写本から抽出できるものすべてと言って良い．

図 2 を例に挙げれば，

題: 和歌のタイトル

作者名: 和歌の作成者

集付: 他の和歌集の出典

本文: 和歌本文

歌番号: 後の時代に和歌につけられた番号

である．このうち，和歌の本文と歌番号はどの和歌データにも出現するが，題，作者名，集付は和歌によって存在するものと存在しないものがある．

ここで注目すべきなデータとして集付が挙げられ，同一和歌が他の和歌集にも取り上げられていることを示している．もともと和歌集は誰かが詠んだ和歌を集めたものであるため，当然別の和歌集で用いられた和歌自身も記載されている場合がある．つまり集付は，写本の書写者もしくは所持者が勉学のために書き入れたメモである．

このようなメモは集付だけではなく，和歌本文に対して誤った書写に対する訂正や注意書きを行っている**書入**と呼ばれるものも存在する．このようなデータは，和歌研究者が和歌に関する新たな知識発見を行う上では有用なデータとなり得るのは疑いのないところではあるが，市販の和歌データ集などにはこのようなデータは含まれていない．

したがって，和歌研究者に求められている和歌データベースは，こうした市販の和歌データ集などには含まれていないデータを容易に和歌研究者によって追加できる機能を持ったものということになる．

3.3 RDF データ利用の妥当性

現在，データの格納形式として一般的に用いられている形式として表形式や木構造形式がが挙げられる．これは 2 節でも述べたように，それぞれ RDB や XML データベース (XMLDB) で管理される．

複数の一般的なデータ形式が存在する中，RDF データベースを用いる妥当性を評価するため，RDB，XMLDB，RDFDB の各データベースでどのような利点，欠点があるのかをデータの追記および参照の視点から比較検討した．

表 1 追加データ格納の比較表
Table 1 comparative table

項目	RDB	XMLDB	RDFDB
既存項目の要素の追加	○	○	○
新規項目の追加	△	○	○
外部データベースへの参照	×	△	○
和歌同士の関係性の記述	×	×	○

表 1 を見ればわかるように，RDB における新規項目の追加は可能であるが，スキーマの変更が必要となるため変更ごとに頻繁にデータ構造の更新を必要とする．また，XMLDB における外部データベースへの参照に， URL を属性に付与することで外部データベースに対する参照は可能だが，外部データベースに格納されているデータについて直接参照することは不可能である．具体的なデータを参照し，閲覧することが不可能である．

一方，RDFDB は，柔軟性の高さや格納データの拡張性により，新規項目の追加や外部データベースのデータを直接参照することが容易である．また，RDFDB のデータでは新たな研究成果で得られた和歌データ内の関係性を記述することが可能といえる．

この結果から，RDFDB は RDB や XMLDB よりも和歌データの格納に有用であるといえる．よって，和歌データを RDF 形式に記述した和歌データベースへの格納を行う．

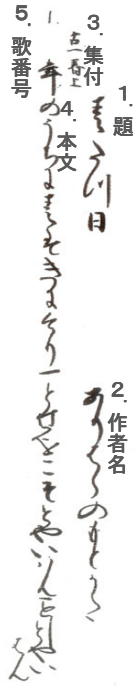

図 2 和歌
Fig. 2 Japanese Poem

4. 和歌データベースの実装

本節では和歌データベースの実装方法について述べる．

まず，和歌を表現するプロパティとして和歌用のプロパティ jpoem[10] を利用し， RDF データの作成を行う．その後， 4store[*2] により RDFDB の実装を行う．

実装に用いる和歌データは和歌研究者から提供された古今和歌六帖である．古今和歌六帖は平安中期に編纂されたと考えられている類題和歌集であり，成立時期や編者はどちらも不明と言われている．収録数は写本により異なるが，各写本には四千数首収録されており，今回は和歌研究者に

[*2] 4store, http://4store.org/

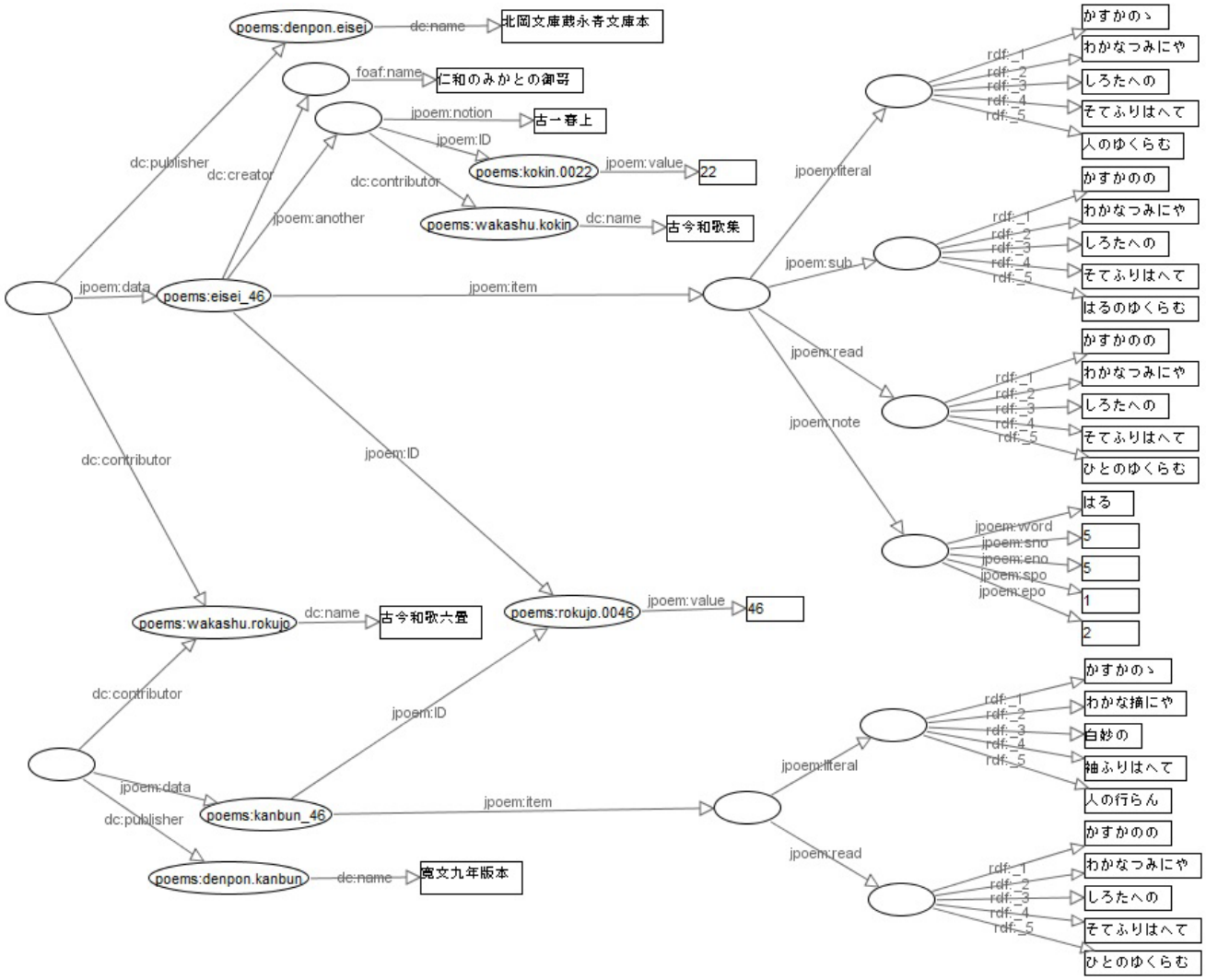

図 3 記述された RDF のグラフの例
Fig. 3 example RDF graph

よってテキスト形式で整形された 6 本の写本を利用する．

4.1 和歌 RDF データ

和歌 RDF データの記述例として古今和歌六帖の歌番号 46 の北岡文庫蔵永青文庫本 (永青文庫本) と寛文九年版本の和歌を図 3 に示す．寛文九年版本の歌番号 46 の和歌には，和歌本文のデータと和歌集に関するデータしか存在しないが，永青文庫本の歌番号 46 の和歌には題，作者名，集付，本文に対する書入の和歌に付与されるデータを保持している．

また， `poems:rokujo.0046` と記述された歌番号を示すリソースを用いて，同一和歌の管理を一元的に行っており，各和歌集ごとに異なる内容が記載されていても，このリソースが示されている和歌は同一和歌であると判断することが出来る．これはつまり，和歌本文表記の異なる同一和歌の検索時においてこのリソースを介して異表記同一和歌のデータを管理している．このとき，`poems` は和歌データを管理している URI の一部を省略した接頭辞であり，本稿で管理されている和歌については全て `poems` が付与される．次に， `jpoem:another` で表現される集付データを利用した他の和歌集への参照データがある．集付データを利用すると，参照可能な和歌集は一つだけに留まらず，異なる和歌集に出典のある同一和歌の参照を行う事が可能となる．これにより，他の和歌集の同一和歌を同時に閲覧することが可能になり，書かれた年代による比較も可能となる．最後に，検索において特に重要となるよみの `jpoem:read` と書入の指示に従って書き換えた本文のよみ `jpoem:sub` がある．よみだけでなく，書入の指示に従って書き換えた本文のよみも対象にすることによって誤字が発生し，通常の語句検索では検索できない和歌についても検索し，抽出することが可能となる．

4.2 4store の利用

4store は RDF クエリエンジンであり，かつデータベー

ススストレージでもある．RDF データを格納した上で，データに対して加工や検索を行える機能を有している．クライアントライブラリが PHP，Ruby，Python，Java と充実していることや，フリーで配布されているため，誰であっても導入しやすい．そこで，本稿でも各自のマシンでデータベースを利用するためのライブラリの選択の幅が広く，他のデータベース利用者が利用しやすい 4store を用いて実装を行う．

検索を行う際に，4store は RDF クエリ言語として SPARQL および SPARQL-Update に対応しているため，トリプルのパターンマッチによって検索やデータの操作を行う事が可能である．

RDF データの記述や RDF データの追加および SPARQL のクエリ式の記述には，RDF の知識や SPARQL の知識といった専門知識が必要となる．しかしながら，提案した和歌データベースを利用する和歌研究者はそれらの専門知識を持ち合わせていない．そのため，専門的な知識のない和歌研究者でも記述可能なようにインタフェースを作成する必要がある．

4.3 和歌データの検索フェーズ

和歌データの検索において既存のデータベースで行える検索を実現することは前提であるが，和歌研究者にとって自由度の高い検索を実現できるように，検索対象および，検索結果の表示パタンを選択可能にすることを目標とする．これにより，和歌研究者の検索に対する要望に応えられるようにする．

具体的な流れを以下に示す．

(1) 検索対象の選択
検索に用いる対象を選択する．その候補として，現段階では和歌集名，歌番号，写本名，作者名，和歌本文の語句，書入により書き換えられた本文が挙げられる．

(2) 検索したい文字列をテキストボックスに入力
検索したいキーワードの検索対象が和歌本文および書入により書き換えられた語句については，検索漏れを防ぐ事を目的としてよみで判断出来るように，検索キーワードをかなで入力する必要がある．

(3) 検索結果の内容を指定
検索結果の範囲を指定する．和歌本文全ての結果を得たいのか，歌番号のみを結果として得たいのかといった，求めたい結果の内容を指定する．結果の候補として，和歌集名，歌番号，写本名，作者名，和歌本文，書入の内容が挙げられる．データが格納されていれば，集付を利用して他の和歌集に出現する和歌の歌番号を指定することも可能である．

(4) データベースに問い合わせて検索
(1) 〜 (3) までの流れを元に RDF データベースに対して SPARQL 式を自動的に組み立て，検索を実行可能となるようにする．検索結果をグラフで記述した場合，膨大なパタンを指定する必要があるため，テキスト形式で結果を表示する．

例えば，図 3 の書入を書き換えた文字列である「はるの」という語句を検索キーワードとし，検索対象を「和歌本文」と「書入により書き換えられた本文」として和歌の検索を行いたいと場合，検索対象を「和歌本文」と「書入により書き換えられた本文」を選択する．次に，テキストボックスへ検索キーワードを入力するときに「はるの」とかなで入力を行い，検索の出力結果の範囲は「和歌本文」と「書入の内容」にする．和歌データベースに問い合わせた際，和歌本文のよみおよび書入のよみのみが検索対象として選択される．この結果，和歌本文と書入のデータが出力される．

このとき，和歌本文に「はるの」という語句は記述されていないが，「書入の内容」の検索結果に「はる」という書入データを含んでおり，「人のゆくらむ」という句を書入の指示により「はるのゆくらむ」に書き換えることが可能であるため，「書入を書き換えた本文」のよみには「はるの」が含まれる．このように，もともとの和歌本文の記述には存在しなかった語句に対して，書入を書き換えたよみの検索も可能となる．

4.4 和歌データの追加フェーズ

データの追加を行う際に，複雑な操作や入力内容を求めてしまった場合，専門知識のない人では操作が困難となってしまう．そのため，和歌研究者が容易かつ直感的に操作可能なように，グラフィカルなシステムを用意する必要がある．以下に，RDF データについて追記を行う場合の構成を記す．

(1) 4store から読み込んだ RDF グラフの表示
4store から RDF グラフを呼び出し，表示を行う．このとき，全ての RDF グラフを記述すると複雑かつ膨大な RDF グラフとなってしまうため，事前に追記したい和歌を指定し，和歌の和歌集名，写本名，歌番号，和歌本文の組を記述した RDF グラフを表示する．

(2) プロパティを付け加えたいリソースを選択
有向グラフで表現された RDF グラフに対して，データそのものやプロパティを追加したい場合に，グラフ上のリソースが記載されているノードを選択する．

(3) 出現する選択ボックスで記述したいプロパティを選択
ノードをクリックすることにより，選択ボックスを出現させる．そこで，記述したいプロパティを選択する．このときのプロパティ候補は書入，作者，題詞，和歌本文の引用，その他の和歌集への参照である．プロパティを記述可能な範囲において選択ボックスにその選択肢が出現する．

(4) プロパティ値に記述する具体的な値を入力

追記するプロパティが決定した後，そのプロパティの持つ値を記述する．このとき，テキスト形式である場合や， URI である場合がある．

(5) RDF データベースに対して追加
(3),(4) で入力されたプロパティとプロパティ値を SPARQL-Update を用いて 4store 内のデータベースを更新する．

(6) RDF データをグラフに新しく反映
(5) で更新した RDF データベースから再び，追加したデータを加えた RDF データを取り出し記述を行う．

例えば，新たに作者のデータを追加したい場合，まず，追加したい和歌データのリソースのノードをクリックすると，選択ボックスが出現し，その和歌データのリソースから派生する可能性のあるプロパティが表示されるが，`dc:creator` を選択する．そして，プロパティを選択したあと，テキストボックスの中にプロパティの値として与えたい値を付与するが，この例であれば，空白ノードにあたるために自動的に空白が選ばれ，さらにプロパティを選択する必要がある．そこで，プロパティ `foaf:name` を記述した後に，テキストボックスに歌人の名前を文字列として入力を行う．この入力が完了した後， SPARQL-Update により，もとの RDF データへ追加を行う．これが反映され，新たに作者データが活かされた RDF グラフを閲覧することが可能となる．

5. おわりに

本稿では和歌研究者らが容易にデータの追加が可能なデータベースの構築およびインタフェースの実装することを目的とし，和歌データのためのプロパティ jpoem を用いた和歌データベースに対してインタフェースの考案を行った．

しかしながら，実際に使用されていないため，和歌研究者にとって使いやすいインタフェースであるかは不明である．

今後の課題として，本当に和歌研究者にとって有用なのかを確認するため，聞き取り調査によるインタフェースの改善とユーザビリティ評価が必要である．

参考文献

[1] ISO 15836:2009: The Dublin Core metadata element set, `http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=52142` (2009).
[2] W3C: Resource Description Framework (RDF) Model and Syntax Specification, `http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/` (1999).
[3] W3C: Extensible Markup Language (XML) 1.0 (Fifth Edition), `http://www.w3.org/TR/xml/` (2006).
[4] W3C: SPARQL Query Language for RDF, `http://www.w3.org/TR/rdf-sparql-query/` (2008).
[5] W3C: SPARQL Update, `http://www.w3.org/Submission/SPARQL-Update/` (2008).
[6] 「新編国歌大観」編集委員会 (編)：新編国歌大観 CD-ROM 版 Ver.2，角川学芸出版 (1996).
[7] 『私家集大成』CD 化委員会（編）：新編私家集大成 CD-ROM 版，エムワイ企画 (2008).
[8] 田中　充，吉村　誠，葛　崎偉：データベース管理システムを用いた万葉集データベースの一構成法，情報処理学会研究報告. 人文科学とコンピュータ研究会報告，Vol. 96, pp. 9–16 (2001).
[9] 竹田正幸，福田智子：古典和歌からの知識発見-モビルスーツを着た国文学者-，情報処理， Vol. 43, No. 9, pp. 941 – 949 (2002).
[10] 白井涼子，波多野賢治：RDF を利用した和歌データベースの構築，平成 24 年度情報処理学会関西支部 支部大会講演論文集，Vol. 2012 (2012).
[11] 白井涼子，欅　惇志，波多野賢治：和歌データの構造化とその格納手法の一考察，電子情報通信学会技術研究報告. DE, データ工学， Vol. 111, No. 76, pp. 79–84 (2011).