RDF を利用した和歌データベースの構築

Structure the Database for Japanese poems Used RDF

白井 涼子 † 波多野 賢治 ‡
Ryoko Shirai Kenji Hatano

1. はじめに

和歌集は古いもので奈良時代から存在するが，刷版技術のない時代には手で書き写されて伝わっている．そのため，同じ和歌集でも漢字とかなという表記の異なりや，単語の表現の異なりが発生する．このとき，書き誤ってしまった場合や書写者があえて表現を変えた場合などが頻繁に起こりうるが，和歌研究においては，その異なり部分に何かしらの特徴があると考え，その時代や言葉の風潮を知見として求め，解明を行っている [1]．たとえば，ある和歌で同じ風景に対して「吉野の里」と表現されている場合と「吉野の山」と表現されている和歌があった場合，時代を遡るほど「山」という表現が増えているのであれば，そこに住む人が増え,「山」から「里」に変わる事が察せられる．これが，表現の変遷に関する知識発見となる．

このような研究を行うに当たり，近年では計算機が使用される場面が増えている．代表的なツールとしては新編国歌大観*や新編私家集大成†が CD-ROM 版 [2] [3] として普及しており，キーワードや歌番号などから短時間でで利用者の検索要求を満たす和歌を検索する事が可能となった．しかしながら，一部の研究者からは上記のような既存の検索が行えるだけでなく，知識発見に役立つ機能を有した和歌データベースを求める声も上がっている．

知識発見へのアプローチとして，データベースの作成は行われていないが，和歌集ごとの傾向を探る頻出文字列を用いた比較分析 [4] や構成文字列の表記が類似した和歌の抽出 [1] などが行われている．同じ和歌集の中でも書き写された複数の和歌集の比較を行うことで，同一部分や異なる部分の差異などを求め，新たな知見を発見することで，和歌研究の成果をあげている [5].

構成文字列の表記が類似した和歌の抽出を行う際，もともと同じ和歌集に収録されていた和歌で表現が変遷していっただけなのか，たまたま似ているだけの全く異なる和歌であるのかが計算機では判断がつかないという問題点がある．このような経緯により，我々はリソースを示すことによりデータの一意性を持つことの出来る Resource Description Framework (RDF)[6] を用いた和歌データのモデルの提案を行った [7]．本稿では，そのモデルをさらに修正した上で，データベースの構築を行う．

2. RDF

本節では，本稿において重要となる RDF について述べる．

RDF は事物の関係性の記述が可能な枠組みであり，有向グラフによりデータ構造が表現され，資源となる二つのノード間の関係性を表現している．その表記形式は複数存在し， N-triple 形式や Turtle があるが，最近よく用いられる代表的な形式として XML 形式で表現されてる RDF/XML 構文 [8] がある．

また，近年ではネットワーク上のデータを参照する場合， Linked Data と呼ばれるデータの表現法があるが，そのモデルの記述にも RDF は利用されている．

2.1 基本的な RDF の構成

RDF では， Subject (主語)， Predicate (述語)， Object(目的語) の三つの要素 (トリプル) で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，トリプルの関係は有効グラフで表すことができる．データ参照の際，RDF では Uniform Resource Identifier (URI) 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る．リソースの場所と名前によって表現されているため，対象となるリソースを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『http://www-ilab.doshisha.ac.jp/ 』，プロパティが『Web サイト』，プロパティの値が 『メディア情報学研究室』となる．この例を日本語で表現した場合は， http://www-ilab.doshisha.ac.jp/ はメディア情報学研究室の Web サイトである，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物体，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．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．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派生する可能性がある．

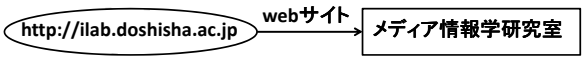

図 1: トリプルの例

2.2 語彙

語彙とは RDF を記述するために必要な表現規則であり，RDF Schema (RDFS)[9] によって記述される．RDFS

† 同志社大学大学院文化情報学研究科，Graduate School of Culture and Information Science, Doshisha University
‡ 同志社大学文化情報学部，Faculty of Culture and Information Science, Doshisha University

*1983 年から 1992 年に書けて刊行された 1162 集の和歌集本文記載した索引

†個人の和歌を編纂した私家集 576 集を纏めた和歌の索引．CD-ROM 版自体は 2008 年に刊行

では，個々のプロパティの定義やプロパティ同士の関係を定義可能である．また，同じ性質のリソースをグループ化するクラスを定義することが可能である．W3C からは，クラスを表現するための基本クラスや基本プロパティといった基本語彙を提供されており，個人が自由に語彙を設計できるようになっている．最近では，必要なクラスやプロパティをすべて最初から作成するのではなく，他の人が作成した語彙が公開されている場合には，それを利用する動きがある．

一般的に，RDFS を記述する際には，語彙によるクラスやプロパティ表現の異なりが発生しないように，まずは既存の語彙で利用できるものがないかを調べ，適した語彙のプロパティが存在場合にのみ語彙を作成する．このとき，一つの語彙で RDF モデルの記述内容を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．

代表的な語彙として Dublin Core[10] が挙げられる．Dublin Core とは Web 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．たとえば，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` という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，さらに詳細な意味を示したい場合には定義域や値域を設定されており，`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．

3. 和歌データ

和歌データとは，和歌本文に加えて和歌集に記載されている作者名や題，書き込まれた文字など和歌に付随しているデータも含まれる．これについては 3.1 節で詳しく述べる．

3.1 和歌研究に用いられる和歌データ

写本に書かれている和歌データの具体例を図 2 に挙げる．記述されている和歌データをまとめると以下のようになる．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．集付とは，同一和歌が他の和歌集に出ていたことを書き記したものである．もともと和歌集は誰かが詠んだ和歌を集めているものであり，別の和歌集で用いられた和歌も記載されている場合がある．これは，写本の書写者や所有者が書き入れたものである．

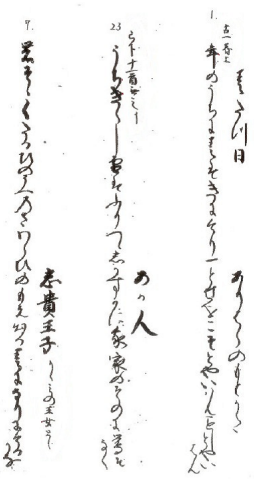

図 2: 和歌データ

3.2 和歌集

和歌集とは，何かしらの題材や人物について和歌を編纂したものであり，天皇や上皇の命により編纂された勅撰和歌集と，個人が撰出した私撰和歌集がある．現存する和歌集は，編纂されたオリジナルである原本から多くの人々の手によって書き写され，現代に伝わってきた．その伝播過程は 3 のような系統樹によって表現することができる．矢印は写本の書写過程を表しているが，現代ではオリジナルの和歌集が残っていることは稀であり，書き写された和歌集である写本が複数存在することが多い．

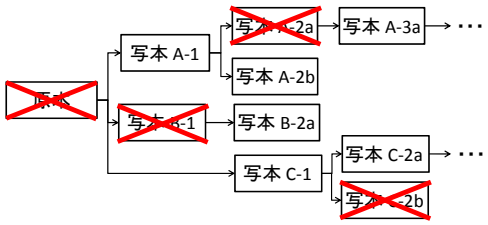

図 3: 和歌集の系統樹

表記が同じ写本は書写過程での写し間違いや，所持者が故意に表現の変更を行うこともあったことから，表記が同じ写本は存在しないと言われている．また，所持者が学習のためや本文の訂正のために，本文の横にメモ書きを残している場合もある．このメモ書きは書入れと呼ばれる．

また，そのような書入れの存在は認識されているものの，和歌研究者に広く普及している新編国歌大観には反映されていない．

3.3 本稿で用いる和歌データ

3.1 節で述べた和歌研究に用いられる和歌データに加えて，3.2 節で述べた和歌集についてのデータを含める．本稿では，本文の訂正やメモ書きとして写本に書き込まれている書入れも特に考慮する．

図 4 に書入れの例を示す．

本文の 1 句目には「やまかせに」とあるが，その上に「谷イ」という書入れがあり，下に書かれた文字と上に書かれた文字を入れ替えるという指示の書入れである．この例では，「やま」の部分を「谷」に変更するという意味を持つ．このとき，書き換え後の読みのデータを生成し，本来の検索では「やま」でなければ検索結果が返されないものを，「たに」でも検索を可能にし，写し間違いであった本文でも検索可能になる．より多くの研究対象になりうる可能性を持った和歌を検索結果として返すた

記述された本文
　谷イ
　やまかせに　とくるこほりの　ひまことに　うちいるなみや　春のはつ花

本文のよみ
　やまかせに　とくるこほりの　ひまことに　うちいるなみや　はるのはつはな

書き換えに対応したよみ
　たにかせに　とくるこほりの　ひまことに　うちいるなみや　はるのはつはな

図 4: 書入れの例

めに，書入れを修正した書き換えに対応したよみも和歌データに含める．加えて，もともとどのような書入れがあったのかというデータも必要と考え，和歌データとして含める．これは書き込まれた「谷イ」という文字と，その書き込まれた位置のデータとする．

以上のことより，利用対象となるデータは 3.1 節の題，作者名，集付，本文，歌番号であるが，本文は実際に記述された様式と，そのよみ，さらに書入れに対応したよみと書入れの表記を含める．

4. RDF のデータモデル

本節では 3.3 節で述べたデータを利用し， RDF におけるデータモデルについて述べる。

4.1 和歌用語彙 jpoem

和歌に対応した語彙を定義する．和歌（Japanese Poem）より，語彙 jpoem を準備する．jpoem はクラスに Poem のみを持ち， Poem の下に複数のプロパティを持つ構造と定義する．

プロパティの種類は表 1 に示し，プロパティの関係図を図 5 に示す.

表 1: jpoem:Poem のプロパティ

プロパティ	中身
data	和歌データ全般
ID	歌番号の ID
value	歌番号の数値
another	集付による参照
nothion	集付の実際の値
item	和歌本文に関するデータ
literal	和歌本文が記述されたデータ
read	和歌本文の読み方のデータ
sub	和歌書き換え後の読み方のデータ
note	書入れの表記に関するデータ
sno	書入れの開始句番号
eno	書入れの終了句番号
spo	開始句番号での書入れの開始位置
epo	終了句番号での書入れの終了位置
word	実際に書き込まれている文字列

まず，jpoem:data が和歌一首を持つものとする．ここで，一元的に管理するためのリソースは和歌集に付与されている歌番号を参照し，プロパティjpoem:ID で示す．さ

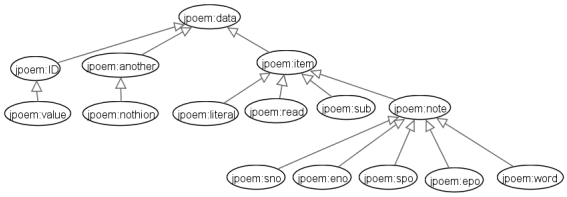

図 5: プロパティの関係図

らに実際の和歌番号の数字をプロパティjpoem:value で示す．集付のリソースはプロパティjpoem:another で示され，実際の表示は文字列としてプロパティjpoem:nothion で示す．和歌の本文の表記はプロパティjpoem:literal によって示され，そのよみはプロパティjpoem:read，書き換えに対応したよみは jpoem:sub で表される．

また，本文に対する書入れについてのデータのを示す．書入れの開始句番号と終了句番号をそれぞれ jpoem:sno と jpoem:eno と表し，開始句番号に対する開始位置と終了句番号に対する終了位置をそれぞれ jpoem:spo と jpoem:epo と記述する．さらに，実際に記述された文字列を示すプロパティを jpoem:word として利用し，それらを統括して書入れのデータ全体を jpoem:note で示す．

その和歌の本文に関する四つのリソースはプロパティ jpoem:item で示す.

4.2 具体的な記述例

記述例として古今和歌六畳という和歌集を用いた．古今和歌六畳には十本程度の写本が現存しているが，そのうち，題，作者名，集付け，のある宮内庁書陵部蔵桂宮本 (桂宮本) の歌番号 1 の和歌と，本文に対する書入れのある北岡文庫蔵永青文庫本 (永青文庫本) の歌番号 5 の和歌を選択した．さらに，同一和歌であることを参照するリソースを表示するため，永青文庫本の歌番号 1 の和歌も例に挙げた．図 6 に RDF のグラフの具体例の図を示す．グラフの記述には，メタデータの記述を行えるソフトウェアの mr3* を用いた．

図 6 において，外向きのリンクしか持たないノードが jpoem のクラス Poem である．これは写本ごとに存在する．このとき， dc:contributor によって参照されているリソースは写本の所蔵を示しており，今回は桂宮本と永青文庫本のリソースを参照し， dc:publisher を用い

*http://mr3.sourceforge.net/ja/

て名前の文字列を示している．次に，jpoem:data によって参照された空白ノードはそれぞれの和歌の和歌データ示している．jpoem:ID で参照されるリソースはどの写本のどの歌番号かを示しており，今回は古今和歌六畳の歌番号 1 と歌番号 5 の和歌に対応しており，jpoem:value によって歌番号の文字列を持つ．この jpoem:ID で示されたリソースが特定の和歌集に付与された歌番号のデータ，rokujo.0001 と rokujo.0005 をそれぞれ参照している．

rokujo.0001 は jpoem:katsura_1 と jpoem:eisei_1 から参照されていることがわかる．これは，桂宮本と永青文庫本の 1 首目に出現する和歌が古今和歌六畳において歌番号 1 の和歌と対応していることが示されている．これにより，歌番号を参考にし，異なる写本でも同一和歌であることを示し，一元的な管理を行うことが可能となったと言える．

和歌本文に対する書入れについては，永青文庫本の歌番号 5 の和歌に着目する．この具体例は図 4 と同じ和歌であるため，テキストとしてのイメージは図 4 を参考にすると良い．jpoem:item で示された和歌本文に関する内容のノードから，jpoem:note によって参照された値が，本文に関する書入れの内容である．この例だと，jpoem:sno が 1 であるため第一句から始まり，jpoem:eno が 1 のため書入れの範囲は第一句のみだと言うことがわかる．さらに，jpoem:spo が 1 であるため，第一句の一文字目から始まり，jpoem:eno が 2 であるため二文字目で終わることが示されている．実際に書き込まれた内容は jpoem:word で示されて,「谷イ」という文字があるということがわかる．

5. 実装

本節では，3.3 節で紹介したモデルのデータベースへの格納を行うための実装に用いる Jena* と，検索に利用する SPARQL について述べる．

5.1 Jena

Jena とは， Apache プロジェクトとして承認されている RDF を扱うために必要な RDF ストレージの一つである．オープンソースのセマンティック Web ツールキットであり，セマンティック Web アプリケーションを構築するための Java のフレームワークでもある．セマンティック Web と Linked Data の開発をのための Java のライブラリとツールのコレクションを提供している．

Jena では以下のことを実現している．

- OWL と FPAFE オントロジーの円環のためのオントロジー API
- RDF と OWL データツールの推論するためのルールベースの推論エンジン
- 大量な TDF トリプルの効率の良いディスク格納が可能
- 最新の SPARQL 仕様書に準拠したクエリエンジン
- RDF データを SPARQL を含む様々なプロトコルを用いた他のアプリケーションを公開

*texttthttp://jena.pabacje.org

5.2 SPARQL

SPARQL Protocol and RDF Query language (SPARQL)[11] は RDF クエリ言語の一種である．RDF で記述されたデータに対して検索を行える言語であり，クエリパターンとして論理積、論理和，そのほかのパターンを指定可能である．記述形式は SQL に近く， WHERE 文も用いられている．Web 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．

SPARQL クエリの例を表 2 に挙げる．

表 2: SPARQL クエリの例

```
PREFIX foaf:  <http://wmlns.com/foaf/0.1>.
SELECT ?mbox .
FROM <members.rdf>.
WHERE{
 ?contributor foaf::name "Ryoko Shirai".
 ?contributor foaf:  ?mbox .
 }
```

まず，PREFIX によって，クエリで用いる語彙を指定し，URL を記述する．検索結果で返してもらいたい値を SELECT で指定する．このクエリの例であれば，mbox は foaf に含まれるプロパティでインターネットのメールボックスを指す．このクエリでは，何かしらのメールボックスの値を返す．FROM では参照する RDF データベースを指定する．WHERE が条件となる．今回は，”Ryoko Shirai” という名前を持つ誰かという条件と，メールボックスを持つ誰かという AND 条件に合致する人のメールボックスを結果として提示するというクエリになる．

5.3 データベースの管理

Jena を用いてデータベースの構築や管理を行う．Jena は Java のではデータの管理についての命令をコマンドライン上で直接入力するのではなく， Java を介して命令を出している．

5.3.1 データの格納

今回のデータは TDB を利用して格納を行う．TDB は RDF ストレージとクエリのための Jena コンポーネントである．Java のプログラムから TDB を呼び出し，格納を行なった．

5.3.2 データの検索

RDF データの検索時には SPARQL を利用する．

通常， SELECT 文でどういった結果が欲しいかを指定するが， SELECT 文では値そのものしか結果として提示できないという欠点があり，求めたい和歌に対して関連するデータを含めて検索結果を提示した場合には利用できない．そこで，CONSTRUCT 文を利用する．CONSTRUCT 文はグラフ・テンプレートで指定されたひとつの結果を XML 形式で結果が与えられるため，結果からグラフを記述する事が可能となる．

本稿では，和歌研究において一意性のあるリソースが重要であることから，二つ以上の和歌を抜き出した場合に，リソースが共通していることを示す必要がある．ま

た，類似した和歌や似ている傾向を持った和歌を比較するために目的の和歌の本文を抜き出す必要も考えられる．

そこで，ある語句をクエリとして検索を行ったとき，その語句を含むという条件に合った和歌の本文と，同一和歌を一元的に管理出来るリソースを提示したいとする．たとえば,「春は」という後で始まる句を必要とした場合に期待する検索結果のグラフを図 7 に示す．

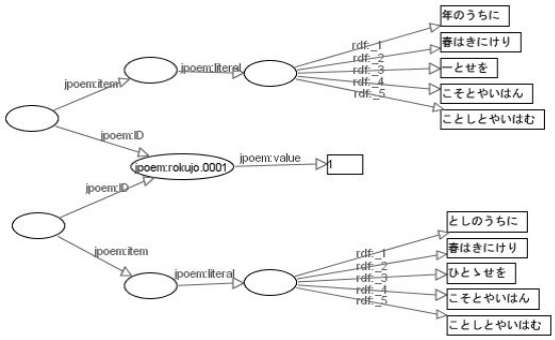

図 7: 期待する検索結果のグラフ

それらを踏まえた SPARQL クエリを表 3 に示す．

表 3: SPARQL のクエリ

```
PREFIX  rdf:  <http://www.w3.org/1999/02/22-
rdf-syntax-ns# >.
PREFIX  jpoem:  < http://sample.org/jpoem/>.
PREFIX  dc:  < http://purl.org/dc/elements/
P 1.1/>.
CONSTRUCT{
  ?about  jpoem:ID  ?value .
  ?about  dc:creator  ?item .
  ?item  jpoem:main  ?literal .
  ?literal  rdf:Seq  ?li .
}
WHERE{
  ?about jpoem:ID ?value.
  {?read rdf:Seq "はるは".}
 UNION {?sub rdf:Seq "はるは"}.
}
```

まず，PREFIX では rdf，dc，jpoem，それぞれの語彙を指定し，URL を記述する．次に，CONSTRUCT で結果として表示したいグラフの部分を記述する．今回は一元的に管理可能なリソースとそのリソースを参照する和歌のデータ，具体的な和歌本文を辿るグラフを結果として求める事が目的であり，jpoem:Data から記載したいリソースを結んでいくことにより，グラフを結びつかせることが可能である．最後の WHERE 文の条件は，まず何かしらの歌番号の値を持っていることと,「春は」のかな表記である「はるは」を読みの部分である read と，読みを書き換えた文字である sub と比較を行いどちらかが true である場合の条件に合った場合に結果を返す式になっている．

6. まとめ

本稿では，和歌データの一元的な管理を行うことを目的としてデータベースの構築を行った．このとき，RDF の構造について述べた後，データモデルの特に重要な部分について説明をし，モデルの内容について述べた．また，RDF の問い合わせ言語である SPARQL について報告を行い，求めたいデータに対する SPARQL の記述を行えた．和歌の RDF を利用したデータベースの構築に必要な項目に関する説明と，実際の構築の状況について述べた．今後の課題として，和歌研究者とのディスカッションを踏まえて，実用性の高いクエリパターンを増やし，システムの実装を行い，ユーザビリティ調査により問題点を洗い出した後に評価を行いたいと考えている．

参考文献

[1] 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.

[2] 「新編国歌大観」編集委員会（編）．CD-ROM 版 新編国歌大観. 角川学芸出版, 1996.

[3] 『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.

[4] 斉藤康彦. 頻出文字列に基づく古今和歌集と新古今和歌集の比較分析の試み. 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. 2009-CH-81, No. 4, pp. 25–32, 2009.

[5] 福田智子. 古典和歌研究における計算機科学の有用性. 村上征勝（編）, 文化情報学入門, 第 3 章. 勉誠出版, 2006.

[6] W3C. RDF. http://www.w3.org/TR/rdf-primer/.

[7] 白井涼子, 波多野賢治. RDF を利用した和歌データの管理に関する提案．情報処理学会研究報告, Vol. 2012, No. 2, pp. 1–6, 2012-03-19.

[8] W3C. RDF/XML Syntax Specification, 2004-02-10. http://www.w3.org/TR/rdf-syntax-grammar/.

[9] W3C. RDF Schema. http://www.w3.org/TR/rdf-schema/.

[10] ISO. The Dublin Core metadata element set. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=52142.

[11] W3C. SPARQL. http://www.w3.org/TR/rdf-sparql-query/.

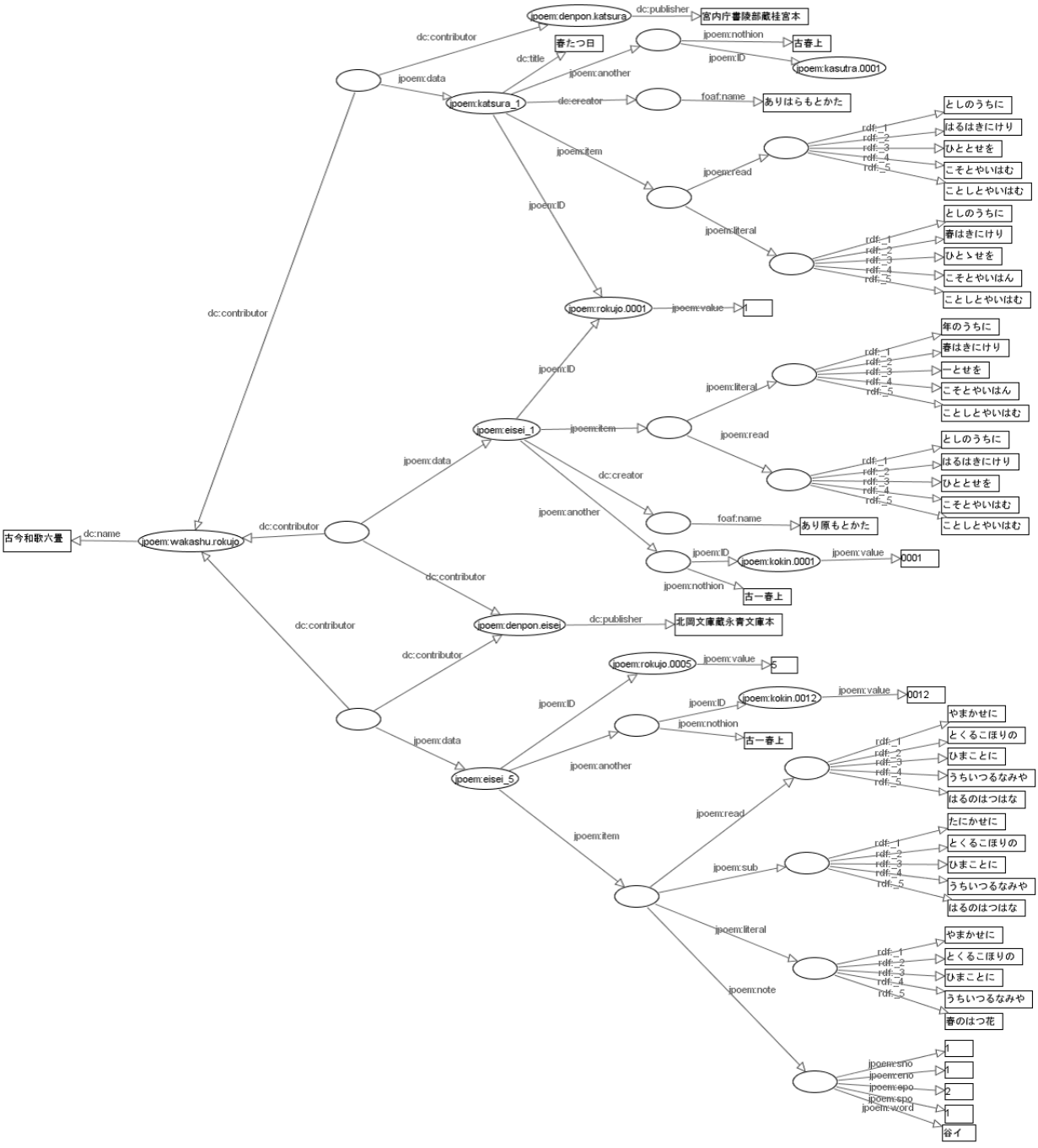

図 6: 記述された RDF のグラフの例