古事類苑の知識グラフ化と言語リソースとしての活用

上松 大輝 [1]　武田 英明 [2,1]　山田 奨治 [3,1]　相田 満 [4]
[1] 総合研究大学院大学　[2] 国立情報学研究所　[3] 国際日本文化研究センター
[4] 日本女子大学
hiroki_u@nii.ac.jp　takeda@nii.ac.jp　shoji@nichibun.ac.jp
aidamitsuru@gmail.com

概要

本研究では，明治期に編纂された百科史料事典である「古事類苑」を知識グラフ化した．さらに，記載された事物やことがら，引用・参照される史料との関係性を元に，古事類苑の言語リソース，およびその基盤として活用方法を提案する．古事類苑は，類書の構造をもとに編纂されており，日本の和歌集や物語といった史料群を体系的に知ることのできる書物を目指して作成された．そこで，古事類苑を知識グラフ化し，さまざまな引用書と語の関係や，現代語との接続を行うことで，明治期に設定された語の分類，および古代から江戸時代までの書物との関連性を扱うことで，言語リソースとして古事類苑を利用可能とする．

1　はじめに

本研究では，明治から大正にかけて日本政府主導で編纂された百科史料事典である「古事類苑」と，古事類苑内で引用書，参考資料として記載されている史料を，知識グラフ化することで、過去の事物の意味や記述の背景となるデータの抽出が可能な言語リソースの構築を目指す．相田 [1] は古事類苑を研究対象として有効活用可能とするために，古事類苑のデジタルデータ化手法の策定を行った．方針として，事項の検索に適した百科事典であること，原本の版組を可能な限り再現できるデータベースであることを目指し，古事類苑の全文テキストデータベースの作成を進めている．また，Uematsu ら [2] は，法人データベースや保険医療機関・保険薬局など日本のベースレジストリとなる情報の知識グラフ，さらに地球上での大きな災害のひとつである地震を観測したデータの知識グラフの構築を行なってきた．特に，知識グラフ構築において重要となるデータ構造を検討する際に，オントロジー指向デザインパターンを用いることで，元となるデータの構造をそのまま活用するのではなく，データの意味や活用方法に適したオントロジーを用いてデータ構造を作成することが可能となる．日本最大の百科史料事典である古事類苑には，万葉集や古今和歌集から引用されたことがらや，古事記を参考資料として編纂されたことがらが記載されており，ことがらとさまざまな書物のリンク構造と捉えることができる．一方で，古典籍である書物にはさまざまな写本や版本が存在し，一意に指し示すことが難しい．上松ら [3] は，，写本や版本などのバージョンを考慮した構造を用いて RDF 化することで，和歌集の参照関係を記述するモデルを提案している．

2　古事類苑

古事類苑は，明治政府の一大プロジェクトとして明治 12 年 (1879) に編纂がはじまり，明治 29 年 (1896) から大正 3 年 (1914) と 36 年の歳月をかけて出版された，本文 1,000 巻，和装本で 350 冊，洋装本で 51 冊，総ページ数は 67,000 ページ以上，見出し数は 40,354 項目におよぶ大百科事典である．熊田 [4] によると，編纂者の一人である西村茂樹は，開国後に判明した欧米列強との大きな差を埋められるよう，国力増強のために学力の向上を目指した．また，先人の著述に依拠し，先人に学ぶという類書の形式にこだわり，大規模で正確な百科事典を目指すため，古代から江戸時代までの書物からの抜粋を用いて国家事業として古事類苑の編纂が始まった．

完成した古事類苑は，西洋型百科全書と伝統的東洋型類書との折衷型となっており，30 部にわたる通読可能な本編と，索引部とで構成されている．本編は，「部」「編」「条」「項」の順に階層化されており，各階層のことがらを端的に表す見出し語が付けられている．さらに，各部の冒頭や各編のはじめにもその内容をまとめた解説が書かれている．また，各

条，項以下には前近代の古典籍から抜粋された引用文が羅列されており，出典となる引用書や解説に用いた参考資料も記載されている．

古事類苑に記載された前近代の知識を活用するために，国文学研究資料館 (国文研) では，全引用書と部分抄分データを，国際日本文化研究センター (日文研) では，索引のデータベースと異体字機能を持つ全文検索データベースが公開されている．これらのデータベースを用いることで，百科事典として索引から語を引き，本編から通読することが可能である．しかし，現在のデータベースでは古事類苑の記述や出典を活用した『古事類苑』そのもの分析や，編纂過程の分析等には至っていない．そこで，古事類苑の部や門の見出し語の階層構造や，引用書・参考資料と見出し語の関係性，出典と部や門ごとの関係性を分析可能とするために，古事類苑のデータ構造に適したデータモデルを作成し，構造化されたデータを作成する必要がある．また，出典となっている和歌集や伝記といった書物の原本の発見とリンクの作成も必要である．さらに，明治，大正期に編纂された百科事典であるため，利用されている漢字にも課題があり，日文研の全文検索データベースでは外字画像で対応しているため，Unicode への置き換えや RDF での外字表現の検討も必要である．

2.1 古事類苑のデータモデル

古事類苑が持つ「部」「門」の見出しと「項」「目」「細目」，さらに引用書，参考資料をもとにデータモデルを作成した．図 1 に古事類苑のデータモデルをグラフ構造を用いて示す．

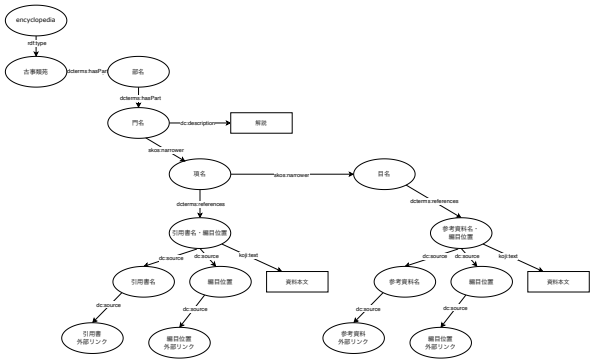

図 1 古事類苑のデータモデル

「部」と「門」は見出しのため dcterms:hasPart を用いて接続され，「項」や「目」は門名に記載された項目をより詳細に記述したものとなる．項目によって，項まで，目までしか持たないものが存在し，各項目として収録された語を利用している文献を示す引用書や参考資料がリンクされる．

古事類苑は，項目が単に並べられた事典ではなく，引用書や参考資料が記載されており，該当の項目が引用書内でどのように利用されているか，また項目の概要を作成するにあたって参考にされた資料を発見することができる．

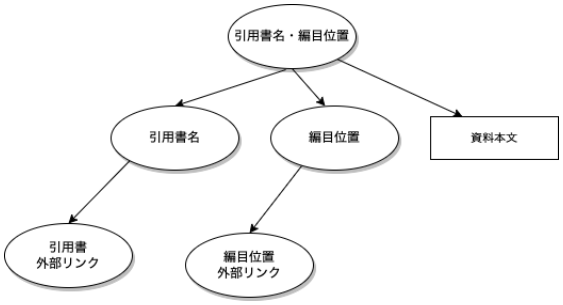

図 2 引用書のデータモデル

図 2 は，図 1 の引用書部分を拡大したものである．引用書や参考資料は，史料の名称と合わせて巻数やページ数など引用箇所や参考とした箇所が含まれる編目位置が一緒に記録されている．ただし，編目位置は史料によって粒度も異なり，編目位置を持たない場合もある．そこで，古事類苑データモデルでは，まず古事類苑記載の史料名称と編目位置を組み合わせたノードを作成した．史料名・編目位置ノードに対して，実際の引用書や編目位置のノードが紐付き，実際に引用や参考にした箇所の本文がリテラルで記録される．さらに，引用書や参考資料本体が，例えば万葉集のように外部公開されたデジタルデータが存在していたり，内部的に編目位置となるページ番号や小番号が定義されていれば，古事類苑知識グラフから外部リソースに対して詳細なリンクを設定することが可能となっている．これにより，部や門ごとに多く引用されている史料や，よく参考にされている章やページなどがわかり，例えば万葉集や古今和歌集では，外部リソースを辿ることで，よく利用されている詠み人を発見することも可能となる．また，古事類苑では利用されている未公開の史料を発見し，新たなアーカイブ作成へとつなげていくことができる，

実際の引用書の例を図 3 に示す．

「神日本磐余彦天皇，〈◯中略〉謂_諸兄及子等_曰，〈◯中略〉聞_於鹽土老翁_『曰，**東**有_美地_，青山四周，」

"天"部の"天"門，"方角"項に記載されている目となる"東"は，日本書紀の第三巻・神武天皇に記載さ

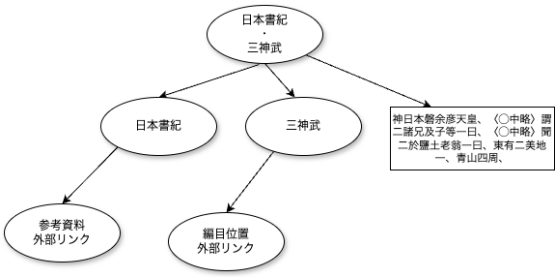

図 3　日本書紀引用時のデータ構造

れている箇所を引用している．

古事類苑は，項目が単に並べられた事典ではなく，引用書や参考資料が記載されており，該当の項目が引用書内でどのように利用されているか，また項目の概要を作成するにあたって参考にされた資料を発見することができる．

2.2　古事類苑の LOD

2.1 節で検討したデータモデルに合わせて，日文研の古事類苑全文データベースを RDF に変換した．古事類苑全文データベースの内部テキストデータには識別タグが付与されており，見出しレベルや解説，引用書や参考資料ごとに識別子が用意されており，

部ごとに門や項のデータを識別子となるタグを行頭に示した後に記載されている．

識別タグをもとに RDF 化し，前後の行に現れる項目同士の関係を抽出しながらリンク構造を RDF で記述し，再帰的に部単位で RDF 化を行うプログラムを作成し LOD へと変換した．例えば，天部の方角に含まれる「東」の URI は以下の構造となる．

http://kojiruien.kgraph.jp/collection/天部/方角/東

各部を古事類苑の collection とし，部の下に門，項，さらに目と続き，URI 上でも事物やことがらの構造を表現している．

3　知識グラフの活用

作成した古事類苑（デジタル化済みの 13 部）知識グラフの詳細と活用事例を述べる．表 1 は，古事類苑に含まれる天部，歳時部，地部，帝王部，封禄部，稱量部，方技部，人部，姓名部，飲食部，器用部，動物部，植物部の 13 の部から作成した RDF の統計データである

対象とした部は全部で 13 部であるが，紙媒体である古事類苑では，巻ごとに部が分割され番号が振られているため，部の数は書籍としての巻の数となり，13 以上となっている．例えば天部には，天という 1 つの部が天部一から天部四，さらに天や方角，日といった 19 の門があり，各門に項や目が続く．また，天部内では，288 の引用書と 33 の参考資料が使用されており，項や目に記載されたことがらに対して引用書名と編目位置，引用箇所の抜粋と，さらに参考資料名と編目位置，参考にした箇所の抜粋が記録されている．引用書と参考資料の被リンク数が，各項や目からリンクされている数を示しており，史料の重複を除くと，引用書が 1800，参考資料が 112 項目記載されていることになる．また，引用書と参考資料で用いられている史料は古事記や日本書紀，万葉集など同じものを利用している場合があり，天部では全体で 308 の資料が利用されている．

表 1　古事類苑 (13 部) の統計データ

見出し	項目数	被リンク数
部	271	722
門	451	12017
項	11107	75802
目	370	2513
解説	429	0
引用・参考	22977	34116

表 2　引用・参考文献上位 20 件の史料

順位	史料名	引用数
1	伊呂波字類抄	538
2	倭訓栞	271
3	倭名類聚抄	245
4	和漢三才図会	179
5	類聚名義抄	171
6	延喜式	161
7	夫木和歌抄	149
8	箋注倭名類聚抄	130
9	易林本節用集	129
10	日本書紀	126
11	続日本紀	122
12	万葉集	120
13	源氏物語	118
14	三代実録	111
15	饅頭屋本節用集	106
16	栄花物語	102
17	一代要記	100
18	重修本草綱目啓蒙	96
19	扶桑略記	96
20	寛永諸家系図伝	93

表 2 は，古事類苑内で引用・参考史料として使われている史料の引用数上位 20 件である．史料数は古事類苑 13 部全体で 3453 件，ひとつの史料あたりの引用回数は平均 5 回であった．表 1 で示したとおり全体では延数 22977 回の引用・参考がされているが，半数以上の史料は 1，2 回程度引用されるのみで，上位 20 件程度の史料のみが多く使用されている．和歌集や歴史書だけでなく，源氏物語や徒然草といった物語や慶長日記などの日記，史料ごとに記載されている内容に特徴があるため，部や門ごとに引用や参考資料としての利用事例に関連があると考えられる

表 3 引用・参考文献上位 20 件の史料

部名	アイテム数	wikipedia	存在率
天部	230	152	66.1
歳時部	610	337	55.2
地部	1682	821	48.8
帝王部	669	193	28.8
封禄部	330	46	13.9
称量部	45	9	20.0
方技部	640	174	27.2
人部	942	300	31.8
姓名部	265	61	23.0
飲食部	730	163	22.3
器用部	1051	137	13.0
動物部	971	226	23.3
植物部	1307	396	30.3

表 3 は，部ごとに項や目の名称の重複を削除した語の数，および Wikipedia 内のエンティティの数を示す．Wikipedia のエンティティは，項や目である「日」「方角」といった語から URL を生成し，実際にページが存在するか確認している．また，古事類苑の「二十四気」は「二十四節気」にリダイレクトされており，「二十四節気」のページが存在するため，エンティティとしてカウントしている．アイテム数および，エンティティ数は部ごとに項目数にばらつきがあり，「ひのもと」や「やまと」，「尾張国」といった地名や「宿駅」や「名所」など土地に関することがらが記載された地部のアイテム数が特に多いことがわかる．Wikipedia のエンティティを見ると，天部，歳時部，地部は 5 割〜6 割程度の語が存在しているが，それ以外の部では 2 割前後である．天部や地部には，天体や気候，地名や地形など，現代につながる地名等が記載されているため，Wikipedia のエンティティも多く，歳時部では年中行事や元号などが多いことが影響している．帝王部には天皇や皇室，方技部には，陰陽道や暦，医術に関することがらが記述されており，一部は現代に通じる語が存在している，姓名部や人部には，親戚，身分や役職名など，現代では利用されていない語であったり，飲食部や動物部，植物部においても現代の常用漢字にない漢字や名称が用いられていることが，Wikipedia のエンティティを発見できない理由と考えられる．旧字体の変換や，現代語への読み替えを行うことで，対応するエンティティを増やすことも必要であるが，部単位ではなく，引用書単位での，語が使用される時代の検証を行う必要がある．

4 まとめ

本研究では，明治期に編纂された百科事典である古事類苑を構造化し，知識グラフとして公開した．類書の構造を持った古事類苑を構造化したことで，事物やことがらを記述するにあたって引用，参考にした古代から江戸にわたるさまざまな史料を分類ごとに一覧することが可能となった．また，Wikipedia 等，現代語のリソースへのリンクを付与することで，古事類苑が日本の古来からの史料を一覧するハブとして活用できる．一方で，元データ内に含まれる外字の変換処理や，現代語との意味を考慮した対応など，言語リソースとしての利用価値を高めるための課題は多々ある．特に，過去の外字データは画像としてデジタル化されており，Unicode 上で存在する漢字へとの対応付け処理が行われていない．守岡 [5] の CHISE 漢字構造情報データベースを用いて部首等の漢字のパーツ情報から，外字データの変換する方法を検討していく．また，参照されている史料にはデジタル化されていないものや，底本や校本が定まっておらず，参照関係が明確でないものなど，より多くの古典籍のデジタル化，史料間の関係性の発見が求められる．今後，言語リソースとしての価値を向上させるために，古事類苑知識グラフ内のデータの充実と利便性の向上，活用方法の提案を進める必要がある．また，古事類苑内のさらなる部のデータ作成を進めるとともに，古事類苑から参照されている外部リソースの発見．構造化を目指す．さらに，数値的な統計解析だけではなく，ある資料を引用や参考に用いた理由や，ことがらと引用箇所の必然性など，文化的な背景の研究への貢献も進めていく．

参考文献

[1] 相田満. 『古事類苑』の共有と近代古典学の解析のための基礎的研究. 情報知識学会誌, Vol. 32, No. 2, pp. 208–213, 2022.

[2] Hiroki Uematsu, Phuc Nguyen, and Hideaki Takeda. Design for data structures: Data unification and federation with wikibase. In **2022 IEEE International Conference on Big Data (Big Data)**, pp. 6169–6178. IEEE, Dec 2022.

[3] 大輝上松, 英明武田, 奨治山田, 満相田. 古事類苑と和歌データベースの知識グラフ構築と相互活用. じんもんこん 2024 論文集, 第 2024 巻, pp. 195–202, nov 2024.

[4] 淳美熊田. 三大編纂物, 群書類従, 古事類苑, 国書総目録の出版文化史. 勉誠出版, 2009.

[5] 知彦守岡, 江一郎江渡, 等流苫米地. Chise project. 漢字文献情報処理研究, No. 4, pp. 58–69, 10 2003.

A　参考情報

古事類苑は，「部」「門」「項」「目」「細目」の順で見出しが設定されており，天部，歳時部，人部，飲食部など大きく 30 の部に分けられている．

表 4 に「天部」の一部を示す．

表 4　古事類苑「天部」データ構造の一部

部	門	項	目	細目
天				
	天			
		名稱		
		天降雜物		
		空中有聲		
		雜載		
	方角			
		名稱		
		四方		
			東	
			西	
			南	
			北	

各部の中に，門名と解説があり，さらに部と門に属することがらが項，目，細目の順で記載される．図 4 と図 5 は古事類苑「天部」の写しである．

図 4 左側には，「天」部「天」門の「方角」項の解説が記載されている．「項」である"方角"が見出しとして段落が開始され，上部には「目」となる"名稱"という記載がある．

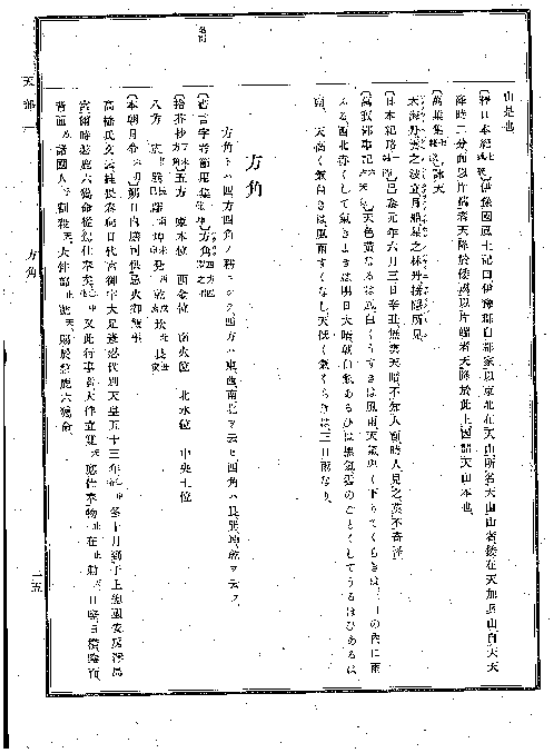

図 4　古事類苑「天部」P.0015

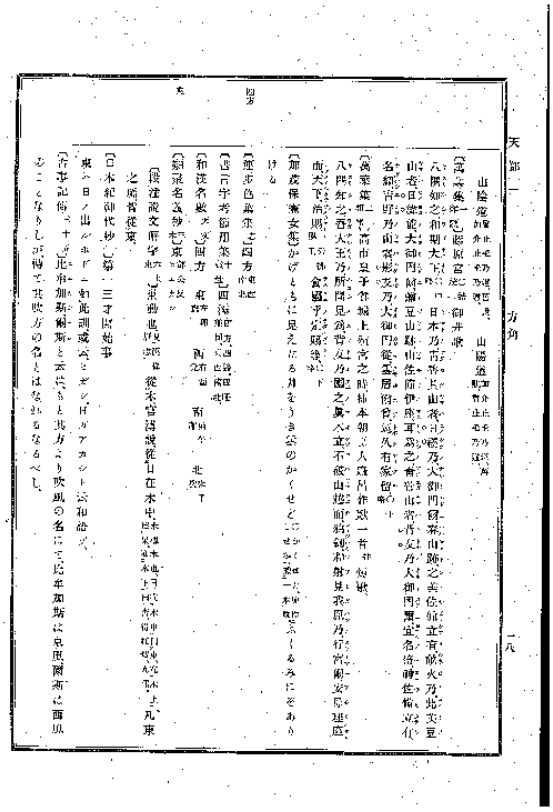

図 5　古事類苑「天部」P.0018