Vol.2019-CH-120 No.10
2019/5/11

# 古辞書をグラフデータベース化する試み —観智院本『類聚名義抄』を例に—

劉冠偉[†1]　池田証壽[†2]

**概要**：平安時代漢字字書総合データベース（HDIC）では，日本の『篆隷万象名義』・『新撰字鏡』，中国の『宋本玉篇』の古辞書データベースを構築・公開した．現在は観智院本『類聚名義抄』の公開を準備している．これらのデータベースは関係データベースであり，FileMaker と MySQL によって管理されている．しかし，研究の深化によって，スキーマの更新やリレーションシップの増加は頻繁に生じるため，関係データベースによる管理は難しくなる．項目・出典・異体字の三者の関係は多・多・多の関係となることが多い．こうした複雑な関係を現在のシステムで対応することが難しくなってきた．直感的なモデリングが可能なシステムが求められている．そこで，本研究では HDIC で公開予定である観智院本『類聚名義抄』テキストデータベースを用いて，収録の各項目，掲出字，注文，所属部首，所在情報を Neo4j に格納して，古辞書のグラフデータベース化を試みる．

**キーワード**：Neo4j，モデリング，平安時代，古字書

# A Graph Database of Dictionary in Early Japan: Based on Case of *Kanchi'inbon Ruiju Myogisho*

GUANWEI LIU[†1]　SHOJU IKEDA[†2]

***Keywords***: Neo4j, Modeling, Heian

## 1. はじめに

グラフデータベースは従来の関係データベースより，高効率で相互依存性が高いデータを処理することができる．古辞書を研究する際に，データベースを構築・利用することが多くなっており，文献の性格を把握するため，原本の構造を“ありのまま”に構築したデータベースが多いが，構築したデータベースを分析するには，コンピュータが理解できる知識の漢字情報が求められる．これらの漢字情報の間には複雑な関係が存在している．

平安時代漢字字書総合データベース（HDIC）[a]では，日本の『篆隷万象名義』・『新撰字鏡』，中国の『宋本玉篇』の古辞書データベースを構築・公開した．現在は観智院本『類聚名義抄』の公開を準備している．これらのデータベースは関係データベースであり，FileMaker と MySQL によって管理されている．古字書の項目は掲出字と注文からなる．原本の構造を再現するため，これらのデータベースは一つの項目を一つのレコードに格納する構造に設計されている．研究の深化によって，スキーマの更新やリレーションシップの増加は頻繁に生じるため，関係データベースによる管理は難しくなる．漢字情報の間，例えば，項目・出典・異体字の三者の関係はグラフ状の多・多・多の関係となることが多い．

こうした複雑な関係を現在のシステムで対応することが難しくなってきた．直感的なモデリングとグラフ状のデータの検索が可能なシステムが求められている．

そこで，本研究では HDIC で公開予定である観智院本『類聚名義抄』テキストデータベースを用いて，収録の各項目，掲出字，注文，所属部首，所在情報を Neo4j[b]に格納して，古辞書のグラフデータベース化を試みる．

## 2. 観智院本『類聚名義抄』データベース

### 2.1 観智院本『類聚名義抄』の概要

観智院本『類聚名義抄』は院政期に成立した漢和辞書『類聚名義抄』の改編本系の唯一の完本として国語学においてよく研究されている．仏・法・僧の三部に分かれ，合計 10 帖からなる．約 32,000 項目に約 42,000 掲出字があり，20 万字以上の注文が収録されている．

一つの項目にある掲出字は，現代の典型的な漢和辞書と同様に 1 字であるものと，熟語または異体字関係である 2 字以上のものが掲出されている．その下にはより小さい文字で掲出字に対する形・音・義を注記する注文がある（図 1）．

項目は同一の大きさで掲載されているが，注文内容によって上位・下位の関係を持っている項目もある．

†1 北海道大学大学院文学院／日本学術振興会特別研究員-DC2
Graduate School of Humanities and Human Sciences, Hokkaido University / JSPS Research Fellowship for Young Scientists
†2 北海道大学大学院文学研究院
Faculty of Humanities and Human Sciences, Hokkaido University

a) https://hdic.jp
b) https://neo4j.com

Vol.2019-CH-120 No.10
2019/5/11

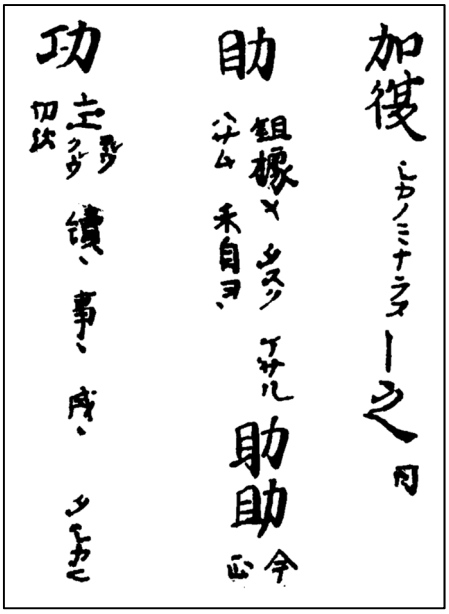


図 1　観智院本『類聚名義抄』記載例（原本模写）

### 2.2 テキストデータベースの構造

観智院本『類聚名義抄』テキストデータベースの構築はすでに終了し，現在は点検・校正の段階である．データベースの構造は次の表 1 のとおり，一つの項目を一つのレコードとして格納している[1].

表 1　データベースの構造

| フィールド | 内容 |
|---|---|
| KR_ID | 掲出字 ID |
| KR2ID | 風間書房版の項目 ID |
| KR_Tenri_p | 八木書店版のページ数 |
| KR_vol_radical | 巻・部首番号 |
| KR_vol_name | 巻名 |
| KR_radical | 部首名 |
| Entry | 掲出字 |
| Entry_original | 原字形に近い掲出字 |
| KR_def | 注文 |
| Remarks | 備考 |

### 2.3 注文テキストの構造化

前述のように，一つの項目は掲出字と注文からなる．注文は，音注・意義注・字体注・和訓の四つの要素に分けられる．さらに，各要素は下位要素を持ち，複雑な体裁で構成されている（図 2）. その詳細は文献[2]を参照してほしい.

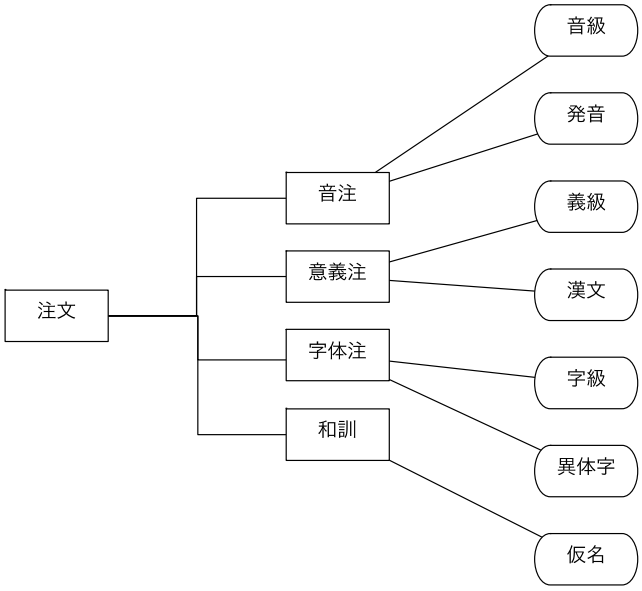



図 2　注文テキストの構造

## 3. グラフデータベース化

### 3.1 目的

漢字の字体情報[c]に関して，グラフで表現できる CHISE プロジェクト[d]があり，その階層的な包摂関係は古文献の文字を翻刻する際に非常に有益である[3]．また，漢字の UCS 符号化組織 IRG[e]の提案漢字を管理するためのグラフデータベースに基づいた文書リポジトリ統合管理システムもあった[4]．グラフデータベースは有用なツールとして人文学分野にも応用されている．

先述のように，観智院本『類聚名義抄』の内容を研究するには，複雑な多対多関係を処理することが必要となる．

例えば，原本の構造〔掲出字―注文〕から抽出できるキーバリュー型データ〔単漢字―和訓〕〔単漢字―異体字〕などの情報が生成され，研究上ではこれらの関係を利用して，〔単漢字①―和訓―単漢字②〕〔単漢字①―異体字―単漢字②〕のように関連漢字を検索することが多く行われている（図 2）.

このような検索は現有の関係データベースで実現することは難しい．高効率で相互依存性が高いデータを処理できるグラフデータベースが目に入ってきた.

c) 漢字の構造，包摂関係，符号位置など
d) http://www.chise.org
e) Ideographic Rapporteur Group

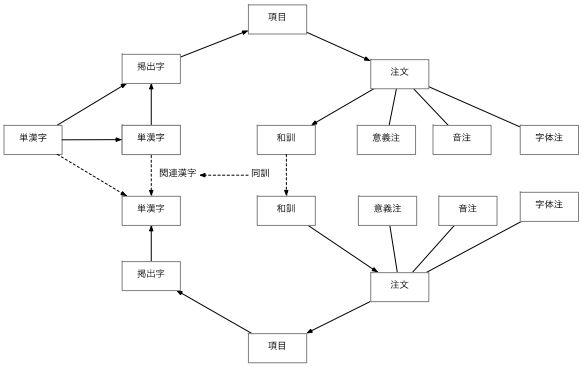



図 2　関連漢字検索のながれ

### 3.2 古辞書のプロパティグラフモデル

#### 3.2.1 プロパティグラフモデル

「物のためのノード，構造のための関係」に準じて[5]，言語学・国語学研究上によく扱われる研究対象をエンティティにする．なお，古辞書における本文内容研究の推進状況により，随時ラベルの追加を行う．

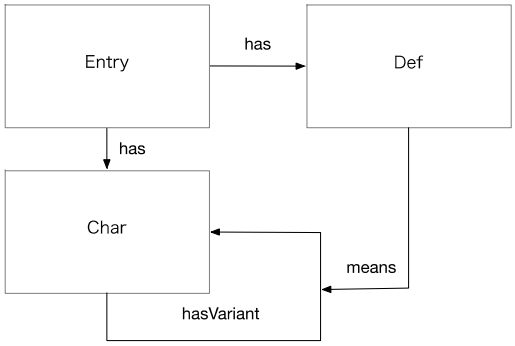
ノードのラベル：項目（Entry），掲出字（Char），注文要素（Def），漢字音（Pron），漢文意義（Meaning），和訓（Wakun）など

関係のラベル:所属（has），解釈（means），異体字（isVariant），「同上」注記[f]（sameAs），注文要素の順序関係（prev, next）など

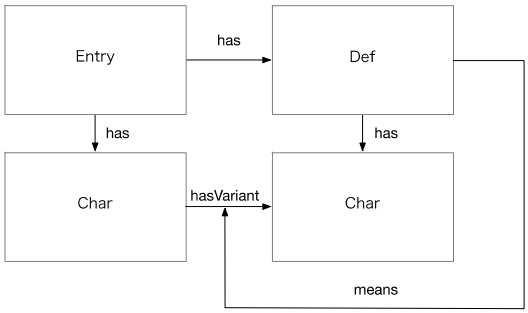
プロパティ：ID，出現順，注文の全文テキストなど

#### 3.2.2 各注文要素のモデル

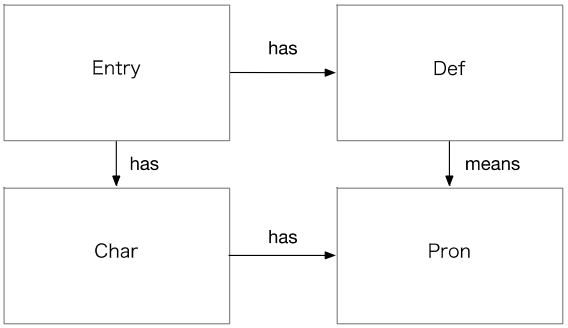
①字体注 A



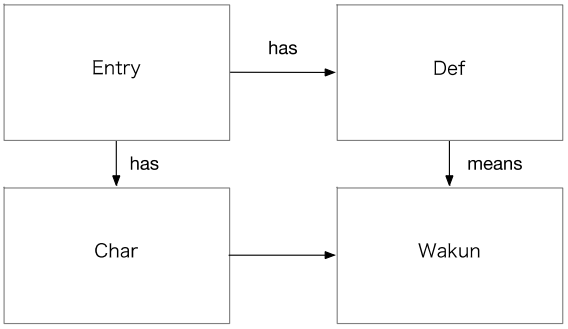

②字体注 B



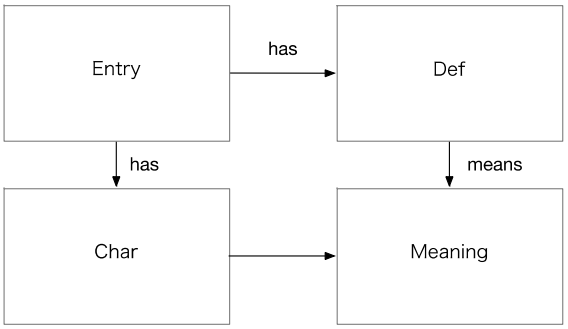

③字音注



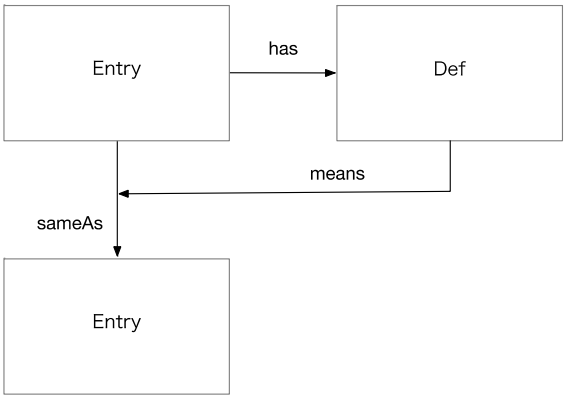

④和訓




⑤意義注




⑥その他（「同上」）




f) 前出の項目と同一であるという注記

## 4. Neo4j による実装

筆頭著者（劉）は日本古辞書における注文構造を効率的にマークアップするためのツール tagzuke を開発して，そのソースコードもすでに GitHub によって公開[g][h]している[6][7]．本研究は tagzuke を用いて，観智院本『類聚名義抄』の注文テキストをマークアップして構造化した．

構造化したデータを加工して，図 3 のような漢字情報を抽出して Neo4j へ導入する．

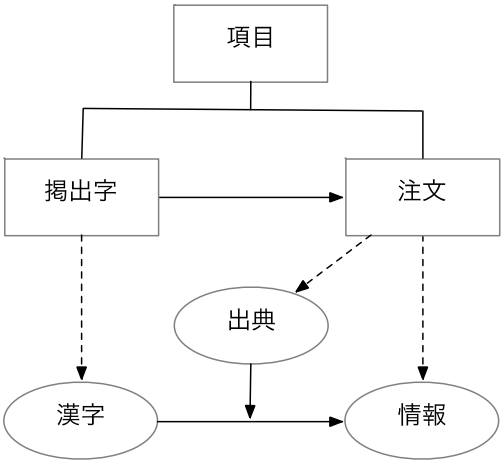



図 3　漢字情報の抽出

## 5. おわりに

本研究では，観智院本『類聚名義抄』テキストデータベースに収録された各項目，掲出字，注文，所属部首，所在情報を用いて，Neo4j での古辞書のグラフデータベース化を作成してみた．

課題として，原文の校正，ネットワーク分析が取り上げられる．

**謝辞**　本研究は JSPS 科研費 16H03422，19H00526 の助成を受けたものである．

## 参考文献

[1] 池田証壽, 劉冠偉, 鄭門鎬, 張馨方. “観智院本『類聚名義抄』全文テキストデータベース構築の方法”．日本語学会 2018 年秋季大会．岐阜, 2018-10-13/14.
[2] 劉冠偉, 李媛, 鄭門鎬, 張馨方, 池田証壽. 部首分類体日本古辞書の項目構造の多様性に対応した マークアップ・ツールの開発．じんもんこん 2017 論文集. 2017, vol. 2017, p. 97‑102.
[3] 守岡知彦. “CHISE のデータ形式 (Ver.0.1)”. http://git.chise.org/~tomo/character/chise-format.pdf, (参照 2019-04-11).
[4] 王一凡, 永崎研宣, 下田正弘. グラフデータベースによる文書リポジトリ統合管理システムの設計. 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. 2018, vol. 2018, no. 8, p. 1‑6. http://id.nii.ac.jp/1001/00187424/.
[5] Ian Robinson, Jim Webber, Emil Eifrem. グラフデータベース. 佐藤直生監訳，木下哲也訳. オライリー・ジャパン, 2015.
[6] 劉冠偉. 日本古辞書マークアップ・ツール tagzuke の課題—操作性・汎用性・維持性の改良—. 情報処理学会, 2018, 1–4p.
[7] 劉冠偉, 李媛, 池田証壽. スマホで古辞書 II—平安時代古辞書の総合的インタフェースについて—. じんもんこん 2018 論文集. 2018, vol. 2018, p. 83–88.

g) https://github.com/toyjack/tagzuke
h) https://github.com/toyjack/tagzuke_cli