社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－CH－72（6）
2006／10／27

古事類苑（天部・地部）の全文入力と Wiki 版の試行
－前近代の文化概念の情報資源化－

山田　奨治 [*1], 早川　聞多 [*1]，相田　満 [*2]

[*1] 人間文化研究機構・国際日本文化研究センター／総合研究大学院大学
[*2] 人間文化研究機構・国文学研究史料館

『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) にかけて出版された，本文 1,000 巻，和装本で 350 冊，洋装本で 51 冊の大百科事典である．そこには，前近代の文化概念について，明治以前のあらゆる文献からの引用が掲載されており，人文科学研究を行ううえでたいへん有用な事典として，いまでも利用されている．この『古事類苑』を電子情報資源化して活用すべく，著者らはその全頁の画像入力を行い，さらに全文テキスト入力作業を進めている．この論文では，現在までの作業の方法・進捗・課題点，インデックスからのシソーラス辞書作成，そして HTML 版の一部試験公開と Wiki 版の試作状況について報告する．

Full Text Input of *Kojiruien (Ten-bu and Chi-bu)* and Wiki Implementation: Making Electric Resource of Pre-modern Japanese Cultural Concepts

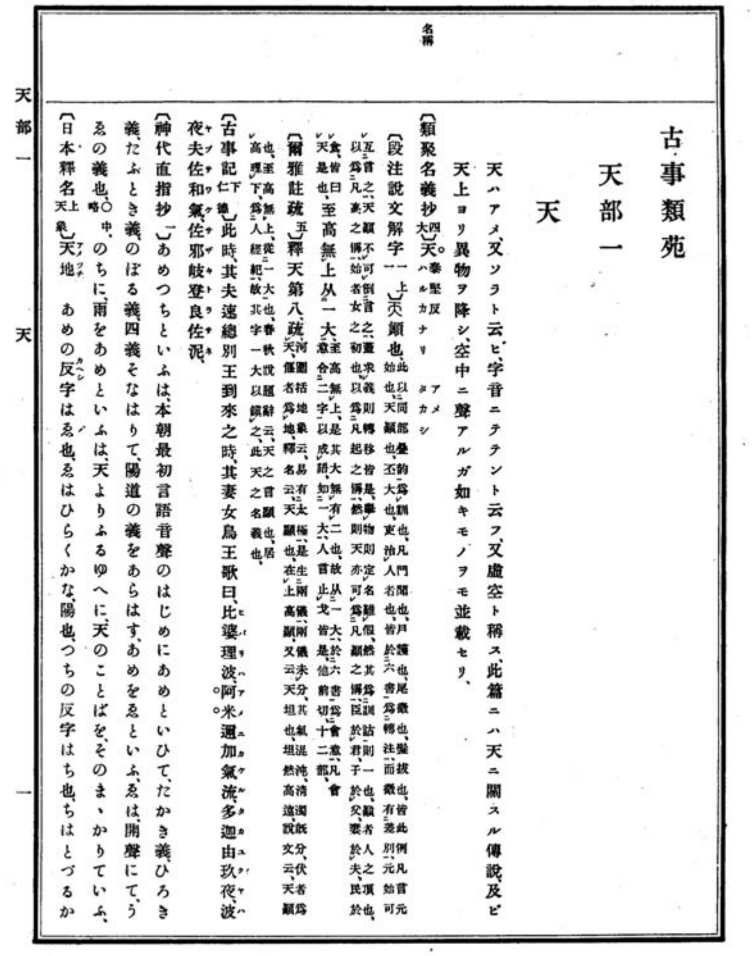
YAMADA Shoji[*1], HAYAKAWA Monta[*1], and AIDA Mitsuru[*2]

[*1]International Research Center for Japanese Studies /
[*2]National Institute of Japanese Literature,
National Institutes for the Humanities
[*1]The Graduate University for Advanced Studies (SOKENDAI)

*Kojiruien (The Dictionary of Historical Terms)*is a kind of large-volume Japanese encyclopedia with 1,000 volumes in original form, 350 volumes in Japanese-style binding, and 51 volumes in Western-style binding, which started editing in 1879 as a big project conducted by Government of Japan, and published during 1896 to 1914. It contains numerous words and cultural concepts of pre-modern Japan, citing plenty of books and documents before Meiji restoration (1867), and is frequently used even today among scholars of the Humanities. We are conducting a digital *Kojiruien* project, with scanning every page images and making full-text data to utilize it as an electric information resource. In this report, we discuss on the project status, method, difficulty, thesaurus extraction from the index, making and public opening of the HTML version, and trial of the Wiki-based system.

1　はじめに

『古事類苑』は，明治期に国家事業として編纂された大百科事典である．その分量は本文 1,000 巻，和装本にして 350 冊，洋装本にして 51 冊，目次は最深で 7 階層・42,188 項目，総頁数は 68,263 頁にも及ぶ．当時の文部大書記官だった西村茂樹 (1828-1902) が，日本独自の百科事典の編纂を建議し，編纂がはじまった．アメリカ元大統領のグラント来日の際，日本の文化についての質問に答えられなかったことをきっかけとなったことが，俗説としてよく知られている [1].


古事類苑

天部一

天

天ハアメ、又ソラト云ヒ、字音ニテテント云フ、又虛空ト稱ス、此篇ニハ天ニ關スル傳說及ビ天上ヨリ異物ヲ降シ、空中ニ聲アルガ如キモノヲモ並載セリ、

〔類聚名義抄 大四〕天。

〔段注說文解字 一上〕天、顚也、

〔爾雅註疏 五〕釋天第八、

〔古事記 下仁德〕此時、其夫速總別王到來之時、其妻女鳥王歌曰、比婆理波、阿米邇加氣流、多迦由玖、夜波夜夫佐和氣、佐邪岐登良佐泥、

〔神代直指抄 一〕あめつちといふは、本朝最初言語音聲のはじめにあめといひて、たかき義、ひろき義、たふとき義、のぼる義、四義そなはりて、陽道の義をあらはす、あめをゑといふ、ゑは、開聲にて、うゑの義也、〇中略 のちに、雨をあめといふは、天よりふるゆへに、天のことばを、そのまゝかりていふ、

〔日本釋名 上天象〕天地 あめの反字はゑ也、ゑはひらくかな、陽也、つちの反字はち也、ちはとづるか

天部一　天　一

図 1: 『古事類苑』天部の版面

『古事類苑』には，前近代に用いられた圧倒的な数の語彙が収録されているばかりか，古事記・日本書紀にまで遡る豊富な引用文献が示されている．これを電子情報化することができれば，人文科学に益するばかりではない．情報科学にとっても，これを知識発見のためのマイニング資源として活用できる可能性がある．

『古事類苑』は，前近代の日本にあった諸概念を天部・歳時部・地部・神祇部・帝王部・官位部など30の部立てに分類し，そこからさらに階層的に諸概念を配置したことに特色がある．つまりことばを50音順に並べるのではなく，一定の指針のもとに概念が整理・分類されている．その分類体系そのものがいわゆるオントロジを形成するシソーラス辞書になっており，古典語彙の言語処理にそれを利用できる可能性がある．

以上のようなことをにらみつつ，われわれは『古事類苑』の電子化作業と試験版の作成・公開を進めている．この論文では，現在までの作業の方法・進捗・課題点，インデックスからのシソーラス辞書作成，そしてHTML版の一部試験公開とWiki版の試作状況について報告する．

2 インデックス，画像入力

『古事類苑』の洋装本第51巻は，部立順の目次と，目次にある語彙を中心に50音順に配置した50音順索引の2種類が収録されている．われわれは，これらのすべての全文テキスト入力を完了した．また索引を除くすべての頁画像のスキャニングも終了している．

2.1 インデックス入力

部立順目次については，国文研の相田が作業を行った．珍しい特殊な文字については，CJK 統合漢字にエクステンション A(6,582 字) が追加された，UCS3.0 の文字セットで可能な限りカバーすることに努めた．その結果，JIS-X0208 で納まらないものが 42,188 件中 715 件あった．そのうち UCS3.0 でカバーできるものが 683 件あり，カバーできないものが 32 件残った．残った 32 件のうち今昔文字鏡番号があるものが 28 件，ないものが 4 件であった．前者については今昔文字鏡番号を，後者についてはその文字を構成する図形要素の情報を入れてある [2].

50 音順索引については，日文研の早川が作業を行った．『古事類苑』の 50 音順索引には，各語彙の冒頭の 3 音しか記載がなく，完全な「よみ」が付されていない．50 音順索引入力にあたっては，完全な「よみ」を付加することを目標に，難読文字については妥当な「よみ」を推定しながら作業を進めた．このようにして作成された索引件数は，64,248 件になった．ただし，50 音順索引の文字セットは UCS3.0 レベルではあるものの，作業進行の都合上，部立順目次で使用した文字との完全な整合性は，現時点では取れていない．

2.2 画像入力

『古事類苑』の頁イメージを画像データ化する作業も，日文研において完了した（図 1）．すべての頁を 300dpi の 256 階調グレー画像でスキャニングし，JPEG 形式で保存している．近い将来，インデックスから選択して全頁画像を閲覧できるようなシステムを構築し，インターネット公開する予定である．

3 全文テキスト入力

3.1 基本フォーマット

『古事類苑』の版型を構成している基本フォーマットの解析は，相田が行った．この事典は西洋型百科全書と中国型類書の融合したものとして構想されたが，傍証として引かれた文献が増殖を重ねた結果，いわば引用の固まりのような構成になっている [3][4]．そのテキストは基本的に，「引用書」と「参考資料」とからなり，それぞれの書名に「編目位置」と「資料本文」が付く恰好になっている（図 2）．

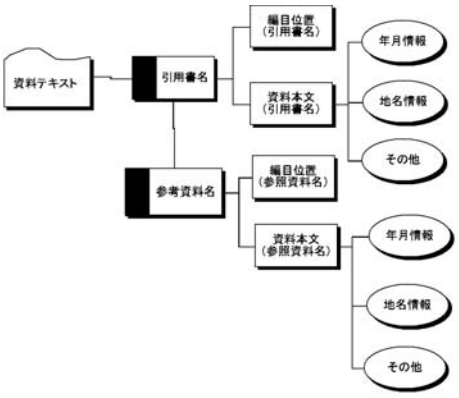


図 2: 『古事類苑』テキスト部分概念図（相田作成，右端の年月・地名・その他は，データベース化の際，マーキングを予定するタグ）

つぎに，版型の基本フォーマットを表 1 のように定めて，全文テキストのタグ付き入力を行った．文字は縦組みで入力し，割注部分は出現位置に〈〜〉で囲ってベタ入力し，割注内改行は無視した．読点，返り点はそのまま入力した．ルビは，(ルS被ルビ文字区1ルビ文字ルE) のようなルビ定型で入力した．ルビは広くアノテーションも含むと解釈し，仮名でなくても文字の右側にある文字はルビ定型で入力した．その際，被ルビ文字とルビ文字の対応関係は，版面上の位置関係よりもむしろ意味の区切りを優先して作成した．文字の左側に記述内容を強調する圏点（○）があるときは，その文字を【〜】で囲んで入力し，○は入力しなかった．

実際の入力作業は凸版印刷株式会社に外注して行なった．文字コードの選択は，版面の「みたまま」を重視し，Unicode でカバーできない文字については，極力，凸版印刷の内部コードを割り当てた．

3.2 校正作業と進捗

以上のように素入力された全文テキスト情報に対して，国文研（相田）と日文研（山田）が校正作業を進めている．両機関とも初校・再校はすべての文字を原本と付き合わせる方法を取り，3 校以降は朱入れ部分のみをチェックすることとしている．

校正で最大の問題になっているのが，特殊な文字に対する文字コードの割付け方である．現在，作業を進行するなかで浮かび上がった問題を整理しつつ，徐々に校正のシステム化を図っているところである．日文研側の作業者が用いている漢字校正ルー

表 1: 『古事類苑』版型の基本フォーマット（相田作成）

	A 部名・付属文字（大字，左右欄外上部）
	B 門名・付属文字（大字，左右欄外下部）
（以下上部欄外）	￥G 概説テキスト
￥C 項・付属文字	￥V〔引用書名 ￥N　編目位置（小字分かち書）〕￥T……○○資料本文○○……
￥D 目・付属文字	￥V〔同（→前項の引用書名に置換）￥N　編目位置（小字分かち書）〕￥T……○○資料本文○○……
￥Q 細目・付属文字	
￥S 細目・付属文字	￥V〔引用書名 ￥N　編目位置（小字分かち書）￥N　編目位置（小字分かち書）〕￥T……○○資料本文○○……
￥T 内訳・付属文字	
	￥Z【イメージ】
	￥W〔参考資料名 ￥M　編目位置（小字分かち書）〕￥X……○○資料本文○○……
	￥P 頁番号（原稿の各頁の開始位置に入力する）
	※以下繰り返し

ルは，図 3 のようなものである．

さて進捗であるが，「天部」（333 頁）は，2006 年 9 月末時点ですでに校了となり，完成されたテクストファイルがある．「歳時部」（1,490 頁）は再校に入っている．「地部（1）」（1,391 頁）は 4 校に入っており，これをもって校了となる予定である．「地部（2）」（1,398 頁）と「同（3）」（1,420 頁）は，素入力を完了し，現在初校段階である．また，まだ入力作業には未着手であるが，「帝王部」（1,690 頁）の一部と「植物部」（総計 2,137 頁）の一部についての入力費用が，別の予算ですでに手当されている．

3.3 Unicode 化と外字

校正作業が終了した「天部」について，全文テキストの Unicode 化と外字作成状況について報告しておく．まず，凸版コードで入力された文字のうち，UCS3.2 への変換テーブルが存在しなかったのは 218 文字だった．これらのうち該当する Unicode が存在しないと認定できた文字数は 21 文字あり，これらについては外字作成が必要となった．もともと凸版コードにもなく，最初から外字作成が必要と認められた 64 文字も含めて，最終的に「天部」について必要な外字は，85 文字となった．ちなみに，過去に国文研において作成されてある外字は 490 文字ある．

表 2: 『古事類苑』天部，地部（1）外字数

	天部	地部（1）
凸版コード数	267	414
うち Unicode 変換テーブルなし	218	384
Unicode 変換不可	21	-
新たに作成した外字数	64	-

一方，「地部（1）」についても同様の統計情報が次第に明確になってきており，それらの概況を表 2 に示した．

これらの凸版コード文字および外字については，フォントイメージの GIF 画像を作成しフォント・サーバを立ち上げて閲覧の用に供することができるよう，凸版印刷側との交渉を進めているところである．

4 シソーラス辞書化と Uniform Concept Locater

さて，前述のように『古事類苑』の目次は前近代の諸概念についての，ひとつの体系的な概念構成を反映したものであるため，これをシソーラス辞書として構築した（表 3）．

このシソーラス辞書の有効性の評価はまだ行って

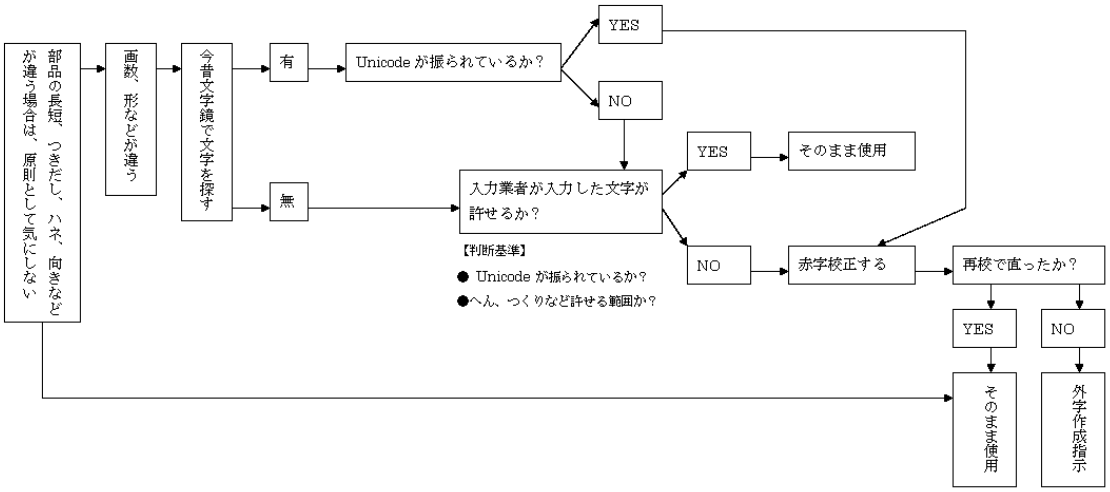


図 3: 『古事類苑』地部入力の漢字校正ルール（山田作成）

表 3: 『古事類苑』シソーラス辞書

語彙	上位語	下位語
root	―	天，歳時，地，神祇，帝王，官位，封禄，政治，法律，泉貨，称量，外交，兵事，武技，方技，宗教，文学，礼式，楽舞，人，姓名，産業，服飾，飲食，器用，遊技，動物，植物，金石
天	root	天，方角，日，月，星，雲，霞，霧，露，霜，雨，雪，霎，霰，雹，風，雷，虹，霽
天	天	名称，天降雑物，空中有声，雑載
名称	天	―
空中有声	天	―
雑載	天	―
方角	天	名称，四方，四角
名称	方角	―
四方	方角	東，西，南，北
四角	方角	艮，巽，坤，乾
東	四方	―
西	四方	―
……	……	……

いないが，前近代日本の諸概念にかんする同種のものがない現状を考えると，こういったものがデータ・マイニングや古典語彙の言語解析の基礎情報として利用されていく可能性があると考える．

また,『古事類苑』オントロジ体系を利用して, 各語彙の概念体系のなかで占める位置を示す，Uniform Concept Locater (UCL) なるものが定義できるのではないかと考える．たとえば,「空中有声」ならば,

UCL://空中有声．天．天．古事類苑/

「東」ならば,

UCL://東．四方．方角．天．古事類苑/

といった記述方式である．

今後インターネット上に全文テキストが公開されていった場合，リソースの概念上の位置を示す方式として，UCL のようなものが必要となってくることも考えられよう．しかし，その評価もまた今後の

表 4: 入力タグから PukiWiki 記法への変換ルール

入力タグ	PukiWiki 記法
A, B	階層化見出しに変換
C, D	*
G	タグ削除
H, X	>
M, N	〈〜〉で囲み，前タグのなかに組み込む
V	**〔〜〕で囲む
W	***〔〜〕で囲む
ルビ	&ruby(ルビ){ 被ルビ文字 };
返り点	&size(5){ 返り点文字 };

（注）P タグは，<page>〜</page> で囲んで，将来的に頁画像とリンクする手段として残した．Z タグは現在使用していない．

研究を待たなければならない．

5 HTML 版と Wiki 版

5.1 天部 HTML 版の試験公開

こうして作成された全文テキスト情報を，広く共有する方法を模索するために，「天部」の HTML 版と「天部」「地部（1）」の Wiki 版を試作した．

HTML 版は，テキストを縦組みで作成し，目次索引と 50 音索引から該当頁にジャンプできるようなインタフェースにした．ただし，検索機能と UTF-16 外の文字にはまだ未対応である．対応ブラウザは Internet Explorer6.0 で，FireFox や Opera では縦組みを表示できない．

この「天部」HTML 版は国文研と日文研のサイトから試験公開している（図 4）[5].

5.2 Wiki 版の試作とタグ変換ルール

われわれは全文テキスト情報の最適な公開方式を探るために，HTML 版のほかに Wiki 版も試作した（図 5）．使用した Wiki は PukiWiki 1.4.7 である [6].

『古事類苑』全文入力時に付けたタグから PukiWiki 記法への変換ルールは，表 4 のとおりである．

Wiki 版を試験してみた結果，(1) 文字コードの問題，(2) ルビ中の返り点処理，が大きな問題であることがわかった．

まずは文字コードの問題である．さまざまな Wiki システムを探索してみたが，残念ながら現在の Wiki がサポートしている文字コードは UTF-8 どまりで，『古事類苑』が求めるレベルには及ばない．これは多くの Wiki が使用している PHP の制約によるもので，この問題はやがて解消されるものという期待を持っている．

つぎにルビ中の返り点処理についてである．現状の PukiWiki のルビ表示プラグインでは，ルビ指示中に文字サイズ変更指示を入れることができないため，ルビ中の返り点を表現することができない．したがって現状では，ルビ中の返り点に限って，【Kレ】のような形で表現することにしている．この問題もまた，プラグインの改良によって解決できる見込みがある．

その他の問題としては，縦組みに対応できない，左ルビを表現できないといった，表現能力の限界が挙げられる．また，実際に Wiki 版が公開された場合，検索要求がサーバにかける負荷が大きな問題となるだろう．

しかし，Wiki 版が持つ優れた共同作業性を生かすことができれば，入力・校正作業を飛躍的に高めることができる．Wiki 版を公開するか否かは未定である．今後各方面からの意見を聴きながら，最適な公開方式を探っていきたい．

6 おわりに

以上，『古事類苑』電子化のプロジェクトの現状について報告した．

最後に課題点として，入力作業をさまざまな機関やあるいは個人が分担する必要性を訴えたい．これまで国文研，日文研の 2 機関において，『古事類苑』の入力作業が進められてきた．その結果，現在までに「天部」「地部（1）」「索引」について全文テキストがほぼ完成し，「歳時部」「地部（2）（3）」「帝王部」「植物部」についても，作業が進行している．これまでの進行状況をみていると，1 機関あたりおよそ 1 年に洋装本 1 冊のペースで作業が進んでいるわけであるが，まったく入力の目処が立っていない残り 44 冊を完了するには，現在のペースを守れたとしても，あと 20 年以上の歳月がかかってしまうことになる．しかし，もし作業を 2 機関ではなく仮に 10 の主体に分散することができたら，あと 5 年程度で全文テキストデータを作り上げることができ

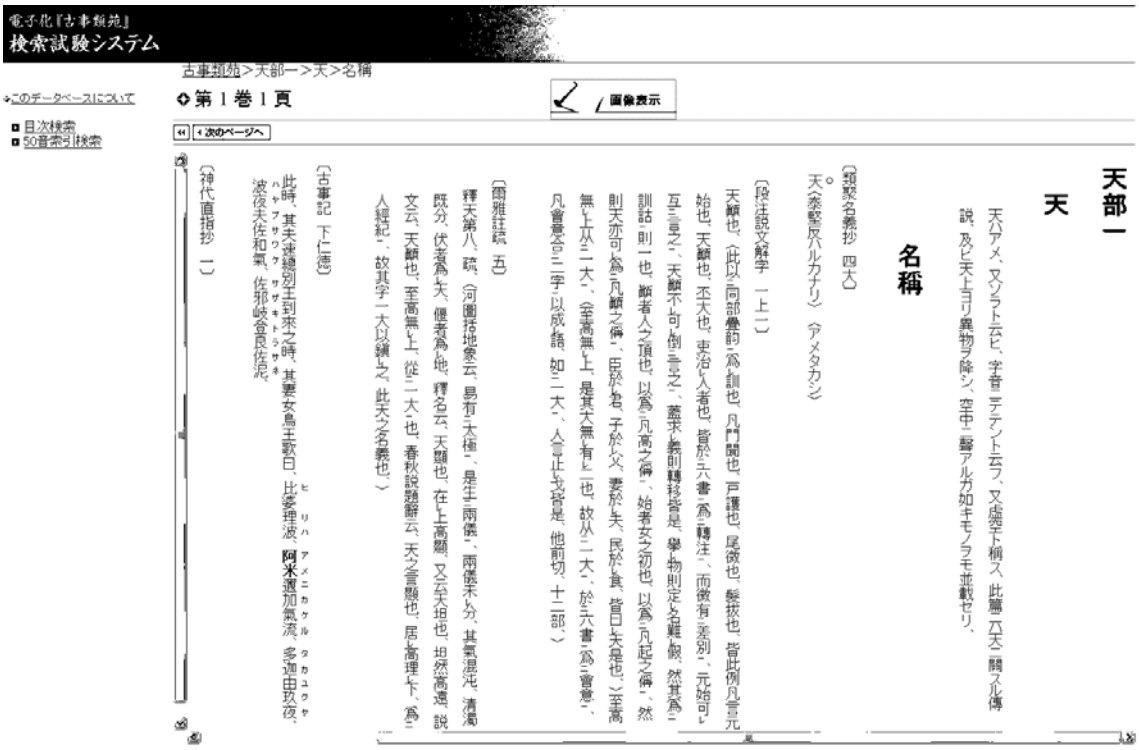


図 4: 『古事類苑』天部 HTML 版

る計算になる．なんといっても時間と費用がかかるのが校正作業の部分であり，そのノウハウの蓄積と作業者の育成には相応の時間がかかる．ただ予算さえあれば，一気に作業が進むという性質のものでもない．さまざまな主体が作業に関わることで，ノウハウを分散共有化し，作業効率を高めていかなければならない．

Wiki 版を公開することができたならば，いま要求されているような共同作業を進展させるために，たいへん有用であろう．

謝辞

この論文の一部は，平成 16〜18 年度日本学術振興会科学研究費補助金・基盤研究 A「前近代の諸概念を対象にした知識発見のためのマイニング資源の開発」(研究代表者 : 山田奨治)，平成 17 年度総合研究大学院大学文化科学研究科「魅力ある大学院教育」イニシアティブ・e-learning 事業群・古事類苑データベース開発事業（代表者：早川聞多)，平成 15〜17 年度日本学術振興会科学研究費補助金・基盤研究 B「和漢古典学のオントロジモデルの構築」(研究代表者：相田満)，平成 17 年度文部科学省科学研究費補助金・特定領域研究「江戸のモノづくり」計画研究「情報技術による知的支援システムの構築」(研究代表者：源田悦夫)，平成 15 年度〜18 年度日本学術振興会科学研究費補助金・基盤研究 B「古典表記構造の統合処理と検索エンジンの研究」(研究代表者：野本忠司）の補助を受けて実施された研究の成果である．

参考文献

[1] 佐藤誠実（著)，瀧川政治郎（編)：律令格式論集, 汲古書院 (1991)

[2] 相田満: 『古事類苑』プロジェクトの構想, 安永尚志（編)：2003 年度総合研究大学院大学共同研究プロジェクト「文化科学研究分野における情報資源共有化のためのコラボレーション研究」第 1 回研究集会報告書, 国文学研究史料館, pp. 170-180 (2005)

[3] 相田満: 日本文化のオントロジー「古事類苑」のデータベース化のために－［論文版］, 和漢古典学のオントロジ 3, 国文学研究史料館, pp. 65-93 (2006)

[4] 相田満: 和漢古典学のオントロジ, 勉誠出版

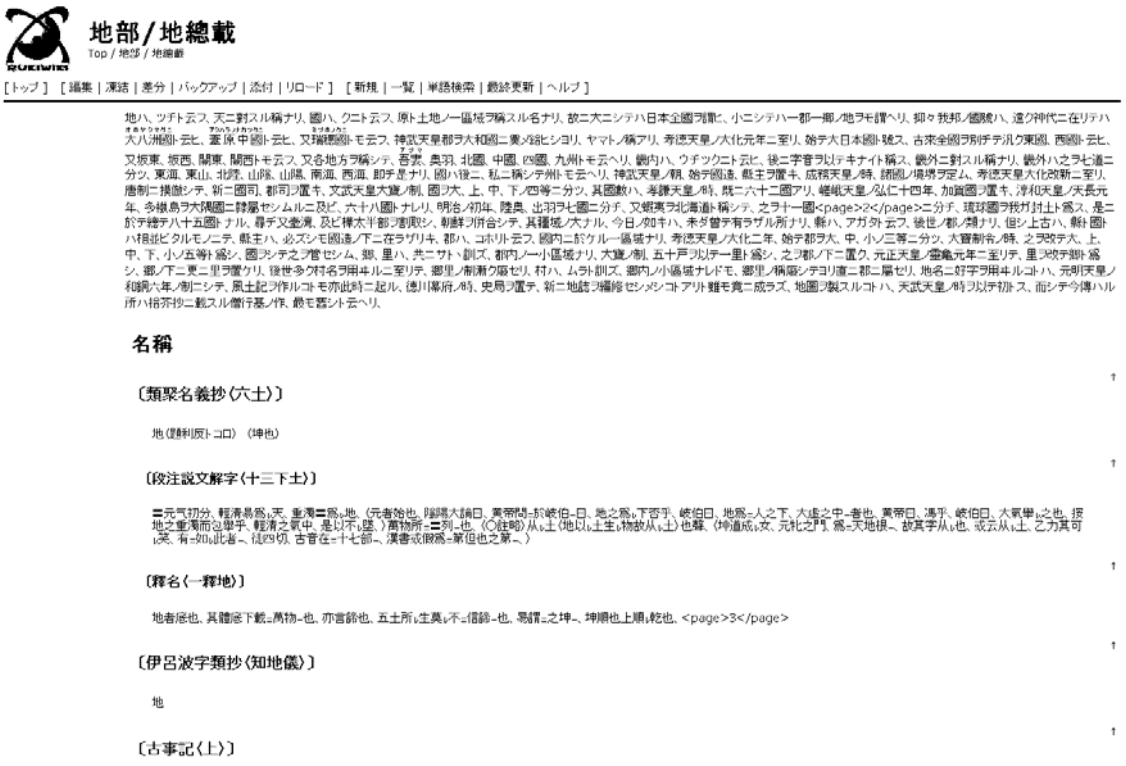


図 5: 『古事類苑』地部（1）Wiki 版

(2007[刊行予定])

[5] 国文研のサイトは，http://base1.nijl.ac.jp/˜kojiruien/ruientop, 日文研のサイトは，http://www.nichibun.ac.jp/˜shoji/kojiruien-test/ten/である．ただし，試験公開版であるので今後予告なく公開内容が変更されることもある．

[6] 増井雄一郎ほか: PukiWiki 入門，翔泳社 (2006)