情 報 学 基 礎　65－1
（２００１．１１．６）

用語体系の統合及び可視化システムの試作

― 用語体系の補完及び相違点の発見を目的として ―

伊東千夏　高久雅生　江草由佳　大懸晶子 *1　宇陀則彦　石塚英弘
図書館情報大学
{chinatsu, masao, yuka, uda, ishizuka}@ulis.ac.jp

シソーラスの統合などの体系の統合が行われているが、概念階層の捉え方は体系により異なっており、統合の問題点となっている。また、体系により階層関係・概念情報など記述されている情報は様々である。そこで、本研究では概念レベルの統合を行わず、どの体系も共通して記述している用語とその用語の階層関係を用い、用語の照合と、階層関係統合の条件の設定により用語体系の統合を行った。語の表記及び語の概念別の階層関係を対象に行ったが、どちらの統合においても、体系の補完と新たな関係の発見に有用であることが分かった。また、用語体系の把握を助ける目的で用語体系可視化システムを構築した。可視化システムでは、ユーザは注目する語やその語の持つ関係を容易に閲覧できるようになっただけでなく、各体系の相違点や統合後の体系の構造を直観的に把握できるようになった。

Integration of Terminological Structure and Visualization System

― as a purpose of complement of terminological structure and finding the difference among source structures ―

Chinatsu Ito, Masao Takaku, Yuka Egusa, Akiko Ohgake,
Norihiko Uda, Hidehiro Ishizuka
University of Library and Information Science
{chinatsu, masao, yuka, uda, ishizuka}@ulis.ac.jp

In this paper, we integrated two or three terminological structures into one structure using matching of words and hierarchical relations. Objects of integration are words distinguished by only spelling, and words distinguished by spelling and concept. It is found that that method is useful in both cases. We also develop visualization system of terminological structures to support understanding of structures. In this system, users do not only browse a term and its hierarchical relationship, but also understand differences among each structure and integrated structures directly.

*1 現在会社員

1　はじめに

概念間の関係や用語の関係を記述した辞書やシソーラスは以前より情報検索や自然言語処理などに用いられてきた。このようなものには自然言語処理用シソーラスである EDR [1]、日本語語彙体系 [2] や、検索用シソーラスである JICST シソーラス [3]、資料分類を目的とした日本十進分類表 (NDC) [4] などがある。また、大規模に作成されたもの以外に、学協会・専門団体などにより作成されたシソーラスや分類があり、一部は Web 上でデータが公開され [5]、個人によるデータの利用が以前より容易になっている。

シソーラスや分類などに記述されている知識に関する情報を有効に利用するために、知識の共有化が進められている。その 1 つとして多言語シソーラス統合プロジェクト [6] がある。これは、WordNet [7]、Euro WordNet [8]、EDR など各国で作成されているシソーラスを統合し、多言語による 1 つの大きな構造を構築することを目的とした多言語シソーラス統合プロジェクトである。

体系を統合する方法は 1) 概念レベルの統一、2) 単語の意味情報を利用した統合、3) 単語の表層的な情報を利用した統合などが考えられる。体系の厳密な統合を求めるのであれば、概念レベルの統一が望ましい。だが、概念レベルの統一については多言語シソーラス統合プロジェクトにおいて、言語間で異なる概念階層の統一が問題とされている [6]。この概念階層統一の問題は多言語の場合だけに生じるのではなく、同じ言語で作られた体系でも、体系ごとに知識や概念の捉え方や作成方針が異なるために起きる問題であり、知識体系を統合する上での問題の 1 つとなる。また、体系に記述されている情報も語の階層関係のみの体系や、語の概念情報についても記述している体系など様々である。統合していくためには、それぞれの体系に共通の情報で、統合していく必要がある。

そこで、本研究では概念レベルの統一を行わず、各体系が共通してもつ情報を用いて体系の統合を行った。統合に用いる情報には、どの体系も共通して記述している用語とその用語の階層関係を用い、用語の照合と、階層関係統合の条件の設定により用語体系の統合を行った。また、各体系の相違点や統合後の体系の構造理解、相違点の発見、用語体系の把握を助ける用語体系可視化システムを作成した。

2　用語体系の統合

2.1　対象データ

複数の体系を統合した場合の問題点を明らかにするために、性格の異なる 3 種類のデータを用いた。データにより概念情報・用語の指示範囲情報の有無があるため、データに記述されている情報から語を概念別に分けて階層関係を記述していると分かる場合は、その階層関係の情報だけを取り出し、どのように概念が異なるかなどの情報は用いなかった。使用したデータは次の 3 つのデータである。

- EDR 専門用語単語辞書（情報処理）中の名詞・概念体系辞書 （以下 EDR）
  EDR では語は概念を示すもので、表層的なものとしてとらえられている。そのため、EDR では同じ語でも概念が異なる場合は別の語として扱っている。今回は、概念体系辞書に記述されている階層関係と専門用語単語辞書の名詞を使用した。
- 日本十進分類表 新訂 8 版機械可読データファイル [9]（以下 NDC）
  日本十進分類表は図書館資料の分類を目的として作成されている。まず、知識を 9 つの分野に区分し、それを更に分野別に分けている。日本十進分類表では分野別に、番号（標数）とその分野を表す語が付与されている。そのため、付与されている語は分野全体を示す場合と、事象を表している場合がある。標数や語の情報のほかに、分野限定や参照情報などが記述されているが、今回は、標数と語の情報のみを使用した。
- 伊東らにより自動抽出された階層関係データ [11]（以下、SS）
  SS は、EDR 専門用語単語辞書（情報処理）の

名詞単語から、SS-KWEIC 法（Semantically Structured Key Word Element Index in terminological Context） [10] により階層関係を自動抽出したデータである。SS-KWEIC 法は専門用語の造語規則を利用し、語のパターンマッチにより自動的に階層関係を抽出する方法である。例えば、SS-KWEIC 法では「検索」と「情報検索」の関係を、「情報検索」に「検索」がマッチするので、「情報検索」を「検索」の下位関係として抽出する。SS-KWEIC 法は語の表層情報を利用しているので、語は概念別に区別されてない。

2.2 統合の方針

データの統合は 2 つのデータによる統合と、3 つのデータによる統合を行った。用語を概念別に分けて階層関係を記述しているデータと概念別に分けずに階層関係を記述しているデータの 2 種類があるので、2 つのデータによる統合では、用語の表記による階層関係の統合と、用語を概念別に分けた階層関係の統合を行った。統合は、用語の照合を行い、完全一致した用語がもつ直下の下位語の階層関係について行った。階層関係の統合にはいくつか条件を設定した。

2.2.1 2 つのデータによる統合

2 種類ずつデータの用語の照合を行い、完全一致した用語の階層関係を統合した。階層関係の統合の条件を以下の 3 種類に設定した。

1-1) 一致した用語の直下の下位語との階層関係のうち、共通している階層関係のみを統合する
1-2) 一致した用語がもつ階層関係のうち、1-1) の条件を満たす階層関係と継承により階層関係が一致するものを統合する
1-3) 一致した用語がもつ直下の下位語との階層関係を全て統合する

用語 A, B, C, D, E について階層関係が記述されているデータ 1、データ 2（図 1）を例に各条件について説明する。用語 A について階層関係を統合すると次のようになる。

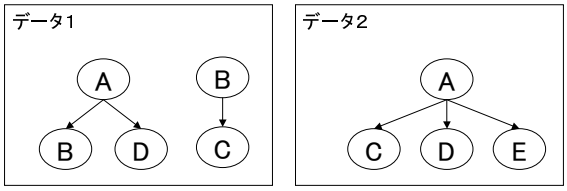
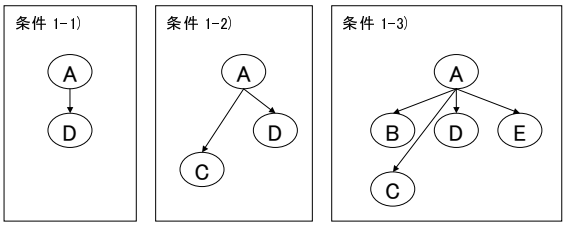


図 1: 2 つのデータ統合例：階層関係の記述



(a) (b) (c)

図 2: 2 つのデータ統合例:統合結果

条件 1-1)

データ 1 は用語 A について A → B, A → D の階層関係をもち、データ 2 では用語 A について A → C, A → D, A → E をもっている。このうち、データ 1, データ 2 に共通している A → D の階層関係のみを統合する（図 2 (a)）。

条件 1-2)

条件 1-1) を満たす階層関係と、用語の階層関係をたどることにより階層関係が一致する関係（継承関係）を統合する。データ 1 では用語 A の階層関係をたどると、A → B, B → C と階層関係がつながっている。この階層関係を 1 つにまとめると A → B → C と記述することができる。このとき、データ 1 の A → B, B → C の階層関係とデータ 2 の A → C とを同じ階層関係と考え、A → C を統合する。条件 1-2) では、継承関係にある A → C と条件 1-1) にあてはまる A → D を統合する（図 2(b)）。

条件 1-3)

用語が一致していれば用語の直下の下位語との階層関係を全て統合するので、データ 1 の用語 A の階層関係 A → B, データ 2 の 用語 A の階層

関係 A → C, A → E とデータ 1, 2 に共通している A → D が統合される（図 2 (c)）。

2.2.2　3 つのデータによる統合

次に、3 種類のデータで用語の照合を行い階層関係の統合を行った。階層関係統合の条件 2-1), 2-3) は先の条件 1-1), 1-3) と同じとし、条件 2-2) を新たに設定し、統合を行った。

2-1) 一致した用語の直下の下位語との階層関係のうち、共通している階層関係のみを統合する
2-2) 基準を定め、直下の下位語との階層関係のうち、その基準を満たした階層関係と 2-1) の条件を満たす関係を統合する
今回は基準を、対象とするデータ数の半数 (今回は 2 つ）が同じ階層関係を持つ場合に統合すると定めた。
2-3) 一致した用語がもつ直下の下位語の階層関係を全て統合する

2.3　統合結果とその考察

表 1: データについて

データ	用語数	概念別用語数	階層関係数	概念別階層関係数
EDR	117,206	117,233	1,799,853	180,381
NDC	30,659	102,639	100,414	102,639
SS	117,206	117,206	119,941	119,941

各データの用語数、概念別用語数、階層関係数、概念別階層関係数を表 1 に示す。

EDR と SS の用語数が同じであるが、これは SS の元となるデータが EDR を利用しているためである。概念別用語数は、用語を概念別に分けて数えたものである。階層関係数は、1 用語がもつ直下の下位語との階層関係の数であり、概念別階層関係数は、用語を概念別に分けたときの 1 用語がもつ直下の下位語との階層関係の数である。SS は用語の情報のみの記述で、概念に関する情報はないが、今回は用語が概念を表していると考え、SS の 1 用語を 1 概念として扱う。

2.3.1　2 つのデータよる統合結果

・用語の表記による階層関係の統合

表 2: 2 データによる統合：用語の表記による階層関係統合結果

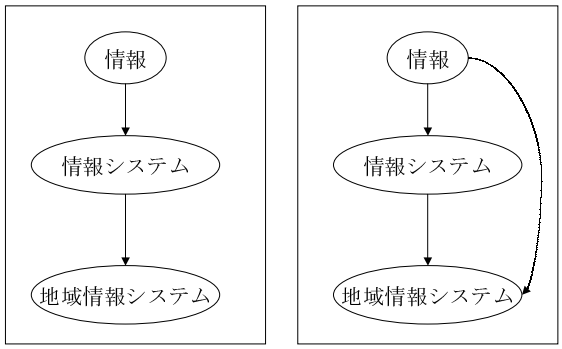
比較データ	一致用語数	条件 1-1)	条件 1-2)	条件 1-3)
EDR, SS	117,206	47,503	14,286	252,291
NDC, EDR	1,166	22	30	25,939
NDC, SS	1,166	59	67	42,731

2 つのデータで、用語の表記により統合を行った時に一致した用語数と一致した用語が各条件における統合でもつ階層関係数を表 2 に示す。NDC と EDR, SS の用語の一致数が少ないが、これは NDC が知識一般を対象にしており、EDR, SS が情報処理分野を対象としているためと考えられる。

条件 1-2) の統合結果から、この統合により体系の中で補完が行われていることがわかった。NDC は、「情報→情報システム」、「情報→地域情報システム」の階層関係をもっており、SS は「情報→情報システム」、「情報システム→地域情報システム」の階層関係をもっている。条件 1-1) による統合では、このうち「情報→情報システム」の関係のみが統合される。統合結果は図 3(a) のようになる。しかし、条件 1-2) による統合では、SS の「情報→情報システム, 情報システム→地域情報システム」が継承関係となり、NDC の「情報→情報システム」と同じ階層関係と考えられ、「情報→地域情報システム」の関係が統合される。統合結果は図 3(b) になり、NDC の「情報→地域情報システム」の階層関係の間には「情報→情報システム」、「情報システム→地域情報システム」という階層関係があるであろう、ということがわかるようになる。

・用語を概念別に分けた階層関係の統合

2 つのデータで、概念別に分けた用語による統合を行った時に一致した用語数と、一致した用語が各条



(a) (b)

図 3: 体系補完の例:情報-情報システム-地域情報システム

表 3: 2 データの統合：用語を概念別に分けた階層関係の統合結果

比較データ	一致用語数	条件 1-1)	条件 1-2)	条件 1-3)
EDR, SS	117,206	47,679	14,462	86,724
NDC, EDR	1,700	23	31	26,198
NDC, SS	1,700	62	70	42,870

件における統合でもつ階層関係数を表 2 に示す。各結果とも、用語のみの結果よりもわずかであるが、一致数が増加している。この増加した関係は、用語による統合では 1 つであった階層関係が、さらに、概念別で分けられたことを示す。

用語を概念別に分けた統合により、用語の表記による統合ではわからない関係を見つけることができた。NDC,SS の用語の表記による統合では、「情報→情報システム」の階層関係は 1 つである。用語を概念別に分けて統合すると、「情報→情報システム」が 2 つ生じる。このことより「情報→情報システム」の階層関係は表記は同じであるが、2 つの異なる概念の階層関係を表しているとわかる。

NDC では、「情報→情報システム」を「007 情報 → 007.3 情報システム (情報と社会, 情報産業)」「007 情報 → 007.4 情報システム」と、「情報→情報システム」を情報と社会のかかわりと計算機による情報システムに分けて記述している。そのため、用語の表記による統合では「情報→情報システム」の階層関係は 1 つであるが、語を概念別に分けて統合すると「情報→情報システム」は階層関係として 2 つになる。このことにより、2 つの「情報→情報システム」は表記が同じであるが、なにか概念が違うものであるということがわかる。

このように、用語を概念別に分けて統合することにより、表記が同じ用語の階層関係であっても、異なる概念の階層関係であるということがわかる。

2.3.2 3 つのデータによる統合結果

表 4: 3 データ比較：階層関係の統合結果

比較データ	一致用語数	条件 2-1)	条件 2-2)	条件 2-3)
EDR,NDC,SS	1,166	14	8,135	54,113

3 つのデータで、用語の表記による統合を行った時に一致した用語数と、一致した用語が各条件における統合でもつ階層関係数を表 4 に示す。3 つのデータによる統合では、階層関係の一致数をもとにした統合 (条件 2-2) を行っている。今回は、2 つのデータで階層関係が一致する場合に統合を行ったが、統合されなかった階層関係の中には「プログラミング言語:論理形プログラム言語」のように、階層関係として正しいと判断されるものが含まれる。条件を定める場合に、階層関係の一致数を用いる他に、対象データの質を考慮した条件も必要であろう。

3 用語体系可視化システムの構築

用語体系を文字として記述したものを眺めるよりも、可視化することにより、各体系の相違点や統合後の体系の構造理解、相違点の発見が容易になるだけでなく、新たな関係の発見などにも有用であると考えられる。そこで、用語体系の把握を助ける目的で用語体系可視化システムを構築した。

3.1 システムの概要

図 4 に本可視化システムの概要を示す。本システムは、各用語体系を記述した XML ファイルを元にインデックスを作成し、ユーザの要求に応じて用語体系の可視化を行い、提示する。

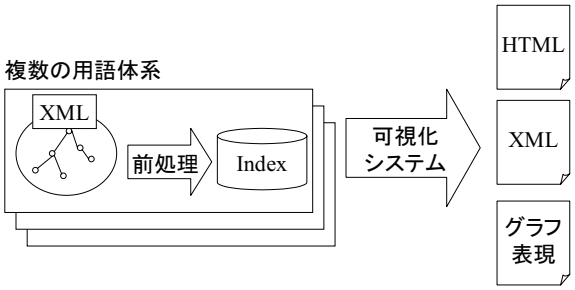


図 4: システムの概要

可視化の表示形式としては、

1. 各用語体系の HTML による一覧表示
2. 各用語体系のグラフ構造のクリッカブルマップ表示
3. 複数用語体系の統合結果のグラフ構造のクリッカブルマップ表示

の 3 つがある。

また、表示したい見出し語を選択するための機能として、

1. 見出し語検索（部分一致、完全一致）
2. 最上位階層の見出し語一覧

がある。

3.2 システムの実行例

以下では、各機能を例とともに説明する。

HTML による一覧表示機能は、各見出し語の上位語と下位語を表示する機能である。表示された上位語と下位語はハイパーリンクになっており、ブラウザ上で語をクリックすることで、指定した見出し語の表示を次々と行うことができる。

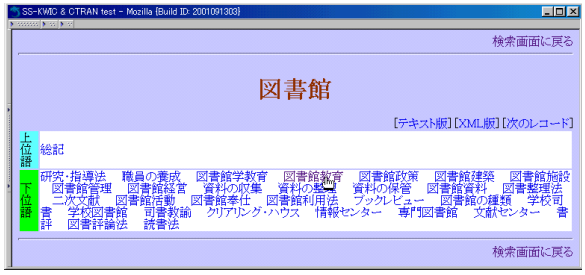


図 5: HTML による一覧表示: 「図書館」（NDC）

図 5 は NDC における用語「図書館」の持つ関係を表示しており、この中から「図書館教育」を選んでいるところである。クリックすると図 6 のように「図書館教育」の上位語と下位語を見ることができる。

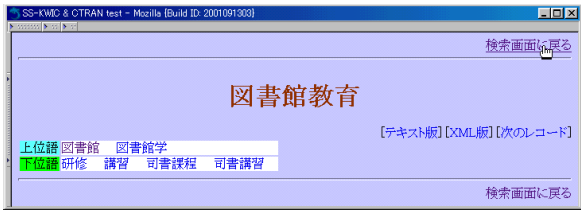


図 6: HTML による一覧表示: 「図書館教育」（NDC）

各用語体系のグラフ構造のクリッカブルマップ表示機能は、各用語体系における見出し語をノードとし、上位階層の語を上に、下位階層の語を下に配置したグラフ表現を表示する機能である。表示された上位語と下位語のノードはリンクになっており、クリックすることで、次々と指定した見出し語の表示を行うことができる。

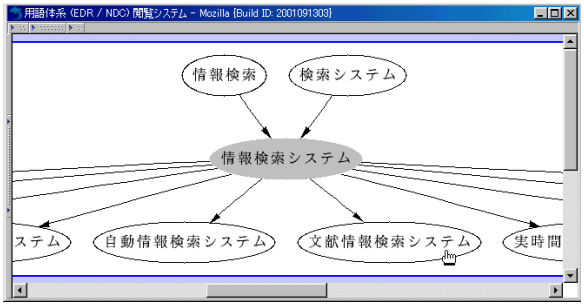


図 7: グラフ構造のクリッカブルマップ表示: 「情報検索システム」（NDC）

図 7 は「情報検索システム」を表示しており、この中から「文献情報検索システム」を選んでいるところである。このノードをクリックすると図 8 のように「文献情報検索システム」についての上位語と下位語を見ることができる。

また、指定した語から上下に何階層まで表示するかも指定できる。図 9 は、NDC の「総記」からたどれる 3 階層分の見出し語を同時に表示した例である。この複数階層の同時表示機能により、可視化する範囲を広げて眺めることが可能になり、ユーザは全体像の把握がより容易となる。

図 10 は、複数の用語体系の統合表示の例である。

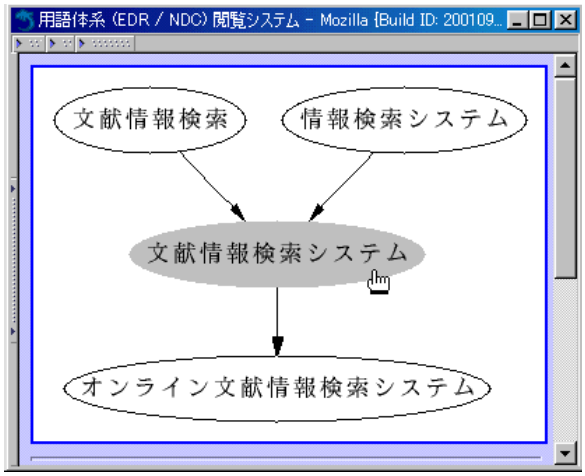


図 8: グラフ構造のクリッカブルマップ表示:「文献情報検索システム」(NDC)

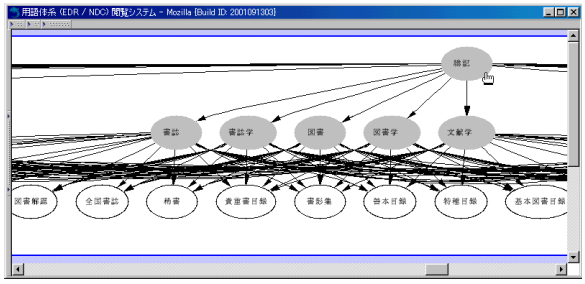


図 9: グラフ構造のクリッカブルマップ表示 (複数階層同時表示):「総記」(NDC)

この例では、用語「情報検索」を例として NDC、EDR、SS の 3 つの用語体系の統合結果に対してのグラフ構造のクリッカブルマップ表示を行っている。複数用語体系の統合結果の可視化では、異なる用語体系の関係が一目でわかるように、SS は赤、NDC は青、EDR は緑というように、各用語体系での関係を表すリンクはそれぞれ別の色として表示している。これにより、各用語体系の間の共通点および相違点が直観的に把握できる。

次に、見出し語検索機能では、各用語体系の見出し語を検索できる。見出し語の検索は、検索したい用語体系を複数選択することで、複数の用語体系を同時に横断的に検索できる。この見出し語検索では、入力された単語と完全一致する見出し語を表示するだけでなく、部分一致した見出し語も表示する。

図 11 は「情報工学」に対する検索結果である。検索結果は、完全一致した語、部分一致した語の順で表

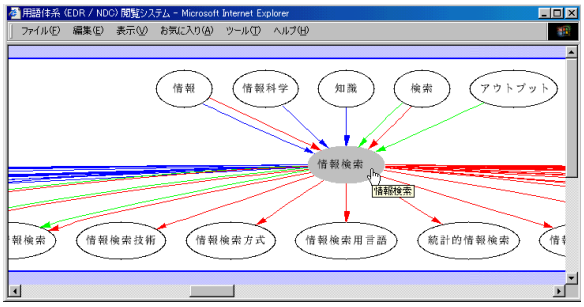


図 10: 複数用語体系の統合結果:「情報検索」

示し、ヒットした見出し語数、用語体系、見出し語の ID、見出し語、見出し語の構造の表示へのリンクが表示される。これらのリンクをたどることで HTML による一覧表示、グラフ構造のクリッカブルマップ表示、統合表示などの表示機能を呼び出すことができる。また、検索結果画面には、すぐに次の検索が行えるように検索結果の下に検索フォームを用意している。

図 11 では、NDC と SS と EDR の 3 つの用語体系すべてで見出し語がヒットしていることがわかる。また、検索語はわかりやすいように赤で強調表示している。この中から見出し語を選択してクリックすると、グラフ構造のクリッカブルマップ表示機能へ飛ぶことができる。また、統合表示欄のリンクをクリックするとその語の統合結果の可視化表示へ、その他の表示形式欄の「HTML」をクリックすると HTML による一覧表示へ飛べる。

図 11: 見出し語検索機能 (検索結果):「情報工学」

3.3 開発環境

本可視化システムは、CGI の枠組みを利用して Web 上でユーザの要求に応じて動的に用語体系を可視化、閲覧できるシステムとして構築した。開発に利用した環境は以下の通りである。Web サーバには Apache Web Server [12] を用い、CGI プログラムは Perl5 で作成した。用語体系の検索などを高速に行うために Berkeley DB ライブラリ [13] を利用した。グラフ構造のレイアウトおよび画像出力は、グラフ描画ソフトウェア Graphviz [14] を日本語に対応するよう改造したもの [15] を用いた。また、XML から各種表示形式への変換には XSLT [16] を用い、XSLT 処理系には Xalan-J [17] を利用した。

4 おわりに

各体系の用語とその階層関係のみを用いて複数の体系の統合を行った。語の表記による統合、語を概念別に分けた階層関係の統合を行ったが、どちらの統合も、体系の補完と新たな関係の発見に有用であることが分かった。

体系を可視化することにより、用語体系の共通点や統合した体系の特徴が直観的に把握できるようになった。また、クリッカブルマップによる表示、表示形式の選択検索の機能を設けることにより、ユーザは注目した語や関係を見れるようになった。

今回は階層関係の統合とその可視化システムの試作を行ったが、今後は、階層関係と同義関係を考慮した体系の統合および可視化システムの構築を行う予定である。

参照文献

[1] 日本電子化辞書研究所. EDR 電子化辞書 2.0 版使用説明書. 東京, 日本電子化辞書研究所, 1999.(TR-006)
URL: <http://www.iijnet.or.jp/edr/>

[2] 池原悟他編. 日本語語彙大系. 東京, 岩波書店, 1997, 5 冊.

[3] 日本科学技術情報センター. JICST 科学技術用語シソーラス　1993 年版. 東京, 日本科学技術情報センター, 1993, 1639p.

[4] 日本図書館協会分類委員会改訂. 日本十進分類法：新訂 8 版. 東京, 日本図書館協会分類委員会, 1978, 635p.

[5] 建築工事標準分類.（参照 2001-10-12）
URL: <http://dbnet.watanabe.arch.waseda.ac.jp/code.html>

[6] 荻野孝野. 異なる言語の繋ぎ手としての多言語シソーラスへの試み人工知能学会誌. Vol.15, No.4, p.567-574(2000 年 7 月)

[7] Fellbaum, Christiane. WordNet: an electronic lexical database. Massachusetts. MIT Press, 1998, 423p.
URL: <http://www.cogsci.princeton.edu/~wn/>

[8] EuroWordNet.（参照 2001-10-12）
URL: <http://www.hum.uva.nl/~ewn/>

[9] 日本図書館協会編. 日本十進分類表新訂 8 版機械可読データファイル (NDC・MRDF8). 日本図書館協会, 東京, 1989

[10] Jinjuan Lai, Hanxiong Chen, Yuzuru Fujiwara. An Information-Base System Based on the Self-Organization of Concepts Represented by Terms. Terminology. Vol.3, No.2, p.313-334(1996)

[11] 伊東千夏, 宇陀則彦, 石塚英弘, 藤原譲. 意味関係抽出手法統合による概念の体系化. 情報知識学会誌. Vol.9, No.4, 2000, p.38-48(2000)

[12] Apache Software Foundation. Apache Project.（参照 2001-10-12）
URL: <http://httpd.apache.org/>

[13] The Sleepycat Software Homepage.（参照 2001-10-12）
URL: <http://www.sleepycat.com/>

[14] AT&T. Graphviz. last update 2000-10-19
URL: <http://www.research.att.com/sw/tools/graphviz/>

[15] 高久雅生. Graphviz-ja. last update 2001-10-11
URL: <http://nile.ulis.ac.jp/~masao/software/graphviz-ja/>

[16] James Clark ed. XSL Transformations (XSLT) Version 1.0. World Wide Web Consortium, 1999. REC-xslt-19991116.
URL: <http://www.w3.org/TR/xslt>

[17] Apache Software Foundation. Xalan-Java.（参照 2001-10-12）
URL: <http://xml.apache.org/xalan-j/>