相互運用可能な用語体系の構築を目指して

— 用語辞書, 分類表, Web ディレクトリを対象として —

Towards an Interoperable Terminological System

— Applications to Term Dictionaries, Taxonomies and Web Directories —

高久雅生*

Masao TAKAKU *

本稿では, 概念間の関係や用語の意味を記述した用語体系として, Web を始めとする情報提供サービスとともに用いられる用語辞書や階層分類からなるディレクトリの構築とその利用について紹介し, 用途に応じてこれらの用語体系を複数のサイト間で相互連携可能とする手法とその展望について述べる.

This paper describes how terminological systems such as Web directories, term dictionaries and taxonomies are built and used. A method of interoperating terminological systems over multiple Web sites along with their own purposes are also described.

1 はじめに

概念間の関係や用語の意味を記述した用語体系は, 情報検索や自然言語処理などの分野で広く利用されている. 近年では, シソーラスや辞書, 分類表, Web ディレクトリなどの様々な用途で用語体系が作成され, 利用に供されている. 例えば, 自然言語処理用辞書として開発された WordNet [1], EDR [2] や日本語語彙体系 [3], 情報検索システム用のシソーラスと

して JICST 科学技術シソーラス [4], 生物医学分野の文献データベースでも使われる MeSH [5], Web 上の情報を網羅する分類体系 Yahoo! [6] や ODP (Open Directory Project) [7], 図書館における図書資料の分類を目的とした日本十進分類表 (NDC) [8] や基本件名標目表 [9] などがある. また, 用語は使われる分野により様々な用法があり, 同一の語が別分野では異なる意味で使われることもあるため, 上に挙げたような大規模な体系以外にも, 分野毎に, 専門機関や研究者・個人などの手による用語集などが作成され, これらも規模や分野を

* 国立情報学研究所
National Institute of Informatics, Japan
masao@nii.ac.jp

問わず公開されている。Web 上では、例えば、環境関係の分野の用語集 [10]、建築工事標準分類 [11]、学術用語集 [12] など、多種多様な用語体系データが数多く公開されており、これら用語体系を利用できる機会が増している。

もともと、これらの用語体系は大量の文献資料の整理や索引作業のために作成されてきた。つまり、大量の情報を整理して提供し、さらにこの組織化された情報を効率的に利用したいという要求から始まったものである。近年では Web の普及とともに、大規模な資料コレクションの提供および利用は珍しいものではなく、用語体系を用いて組織化された情報を入手し活用したいとの要求もますます強くなってきている。また、一般の Web サイトで情報を提供する際にも、分類やキーワードなど一種の用語体系を用いて情報を組織化して提供することが広く行われており、大規模な資料コレクション以外でもその応用範囲は広がっている。このように、用語体系は Web 上における情報提供には欠かせない存在になりつつある。

このように重要性を高めつつある用語体系であるが、それぞれのコンテンツに応じた体系を一から構築し、運用していくとするには多大な労力がかかる。とりわけ、現在の Web のように分散的に多くの場所でコンテンツが作成されているような状況では、それぞれの現場でそのコストを払う

のは、困難な場合も多い。

そこでこの問題を解決するために、既存の用語体系を再利用する手法が提案されている。これは、既存の体系のうち任意の部分を抽出して、新規に作成しようとする体系に取り込んだり、マッピングしたりして、再利用しようというものである。例えば、網羅的な用語体系を作成したい場合には、既存の特定分野向けに作られた構造の一部を参照し、その構造を再利用したり、その逆に、特定分野向けの体系を構築する場合には、既存の汎用的な用語体系を参照し、その構造を周縁領域の構造として再利用したりする手法である。

このような再利用を効率良く行うには、複数の用語体系データ間を連携するための枠組みが必要となる。用語体系は元々、それぞれの組織や分野ごとに独立して作成されており、Web 上で提供される際にも、作成者それぞれが独自の形式で提供している。このため、再利用はそれぞれの用語体系とそのデータ形式に熟知する必要があるため、複数の用語体系をそのまま連携させるのには、労力がかかる。さらに、用語体系データは基礎的な学術情報資源であり、検索システムなど他のシステムとの連携も重要な観点である。そのため、用語体系と他のシステムとを組み合わせる場合には、用語体系自体にアクセスするための統一的なインタフェースが必要になってくる。

本稿では、Web の普及を前提として状況

が大きく変化しつつある現状を整理し、以下の2点について筆者なりの解決手法とその展望を示す。

- 用語体系をどうすれば効率的に再利用できるか？
- 複数の用語体系をどのように連携させるか？

本稿の構成は以下の通りである。2章では「用語体系」についてその呼称や範囲について述べる。3章ではWeb上で提供されている用語体系の例を紹介する。4章では複数の用語体系を再利用し、他のシステムと連携するための手法について紹介する。最後に5章でまとめと今後の展望について述べる。

2 用語体系とは

本稿では、見出し語とその語の持つ関係を記述したものを用語体系と呼ぶ。前節でも述べたが、このような概念のものを言い表すにはその用途に応じて、シソーラスや分類体系、オントロジー、用語集など様々な呼称が使われてきた。しかしながらWeb上などで様々な観点から用語集や分類、シソーラスが作成されている現在では、これらの厳密な区別は困難[13]であり、また、これらの様々な用途のものをまとめて取り扱おうとする場合、元々の用途に限らずまとめて用語体系として位置付けるため、この呼称を用いることとする。

用語体系が持つ関係にも様々なものがあり、例えば、用語間の階層関係つまり上位語と下位語の関係や、包含関係、同義関係、関連関係などである。

具体的な用語体系における関係の記述例を以下に挙げる。

EDR 概念体系辞書では、ある概念に対してその上位概念、下位概念が記述され、各概念の概念見出しに相当する関係が記述されている。例えば、概念ID“2f167c”（情報検索）は、上位概念“2f196f”（アウトプット），“3c7a88”（検索）を持ち、下位概念“3c98a3”（光学的情報検索），“3ca83d”（選択的情報検索），“3cbff9”（文献情報検索）を持つといった関係が記述されている。

十進分類表NDCでは、資料の所属する学問体系全体を10ずつのレベルで細分化し、200（歴史）、210（日本史）、210.6（近代史）、210.61（明治維新）といった10進形式で表現する項目が定義され、各項目に対してそれを表現する見出し語が振られている。

WebディレクトリODPでは、階層構造の各ノードに見出し語を振り、「Top: Computers: Software: Internet: Clients: Chat」といった階層構造の中にWebページを分類していく。

3 Web 上における用語体系の活用

3.1 Web ディレクトリ: Open Directory Project (ODP)

図1 ODP: dmoz.org

Web 上における網羅的な用語体系の代表例である Web ディレクトリは、Web 上に存在するページを組織化し、ディレクトリ状の階層からなる構造に分類したものである。

ここでは、ODP を例にとって説明する(図1)。ODP は全世界のボランティア編集者によって作成され、提供されている Web ディレクトリであり、それぞれの階層は 10～20 程度の下位階層を持つ。2005 年 10 月現在、ODP の最上位階層は、「Arts, Business, Computers, Games, Health, Home, Kids and Teens, News, Recreation, Refer-

図2 ODP における階層: 「コンピュータ: ソフトウェア: インターネット: クライアント」

ence, Regional, Science, Shopping, Society, Sports, World」の 16 の下位階層を持ち、さらに、その各階層の代表的な下位階層にもリンクが分かりやすく張られており、全体のおおよその構造を図示するものとなっている。また、英語だけでなく、各国語に対応したバージョンも用意されており、それぞれの言語によって提供されている Web ページ固有の情報も提供され、また、言語間の階層同士の対応関係も提供されていて、これら各国語版を含め、ODP 全体の階層数は約 66 万階層に達している。利用者はこれらの階層をたどることで、それぞれの分野の代表的な Web ページへのリンクを見つけ、情報を得ることができる。図2は「コンピュータ: ソフトウェア: インターネット: クライアント」という日本語階層を表示した画面である。ここでは、こ

```

<Topic r:id="Top/World/Japanese/コンピュータ/ソフトウェア/インターネット/クライアント">
  <catid>1124401</catid>
  <d:Title>クライアント</d:Title>
  <d:Description>このカテゴリでは、インターネットに関するクライアントソフトに関する情報を
扱うサイトを掲載します。 </d:Description>
  <lastUpdate>2005-07-21 09:07:04</lastUpdate>
  <narrow r:resource="Top/World/Japanese/コンピュータ/ソフトウェア/インターネット/クライ
アント/ブラウザ"/>
  <narrow r:resource="Top/World/Japanese/コンピュータ/ソフトウェア/インターネット/クライ
アント/FTP"/>
  <narrow r:resource="Top/World/Japanese/コンピュータ/ソフトウェア/インターネット/クライ
アント/NTP"/>
  <narrow r:resource="Top/World/Japanese/コンピュータ/ソフトウェア/インターネット/クライ
アント/メーラー"/>
  <narrow r:resource="Top/World/Japanese/コンピュータ/ソフトウェア/インターネット/クライ
アント/ファイル共有"/>
  <narrow r:resource="Top/World/Japanese/コンピュータ/ソフトウェア/インターネット/クライ
アント/コミュニケーション"/>
</Topic>

```

図3 ODP で提供される RDF ダンプ: 日本語階層「コンピュータ: ソフトウェア: イ
ンターネット: クライアント」の抜粋

の階層のさらに下位の階層が列挙されるとともに、この階層に属する Web ページへのリンクが張られている。

ODP では、単に Web 上でディレクトリを提供するだけでなく、そこで作成されているディレクトリデータそのものが提供され、無償で再利用できることが特長の一つである。このデータは XML に基づくメタデータ記述手法の一つである RDF 形式と類似の形式で提供されており、ディレクトリ階層のデータおよびそこに配置された Web ページの情報など、Web サイト上で公開されている全ての情報をデータとして取得し、そのまま再利用することができる。図3に日本語階層「コンピュータ: ソフトウェア: インターネット: クライアン

ト」の抜粋を示す。このデータを使用した再利用の例としては検索エンジン Google がある。Google は自前の検索サイト上に「ディレクトリ」として ODP のデータ全体を使用したカテゴリ情報も提供している。

3.2 環境分野の用語集: EIC ネット

EIC ネット [10] は国立環境研究所が提供する、Web 上の環境分野のポータルサイトである (図4)。EIC ネットにおける用語集では、各用語が見出しとして配置され、これらの用語が環境分野を代表する2階層のトップカテゴリとサブカテゴリの中に組織化されている。EIC ネットのトップカテゴリは、「自然環境, 地球環境, 大気環境, 水・土壌環境, 健康・化学物質, ごみ・リサイクル, エネルギー, エコビジネス,

図4 EIC ネット

環境教育・学習，環境行政，環境一般」の11項目からなり，この各階層に対してそれぞれサブカテゴリが付けられている。このカテゴリ構成の中にニュース記事や用語集を分類し，用語集と環境分野の記事などを同時に提供する工夫を行うことで，環境分野における問題や成果を一般に伝えている。

4 用語体系の再利用と連携

1章で述べたように，用語体系の活用を行うためには既存の用語体系をよりよく再利用し，効率的な情報提供を図るという点を考慮することが望ましい。

用語体系の構築では以下のようなプロセスが必要となる [14]。

- 方針の決定
- 用語の選定
- 構造の決定

- 体系の表現
- 実装
- 利用とその評価

これらのプロセスにおける決定は，用語体系の利用用途に合わせて行われる必要があるが，このうち初期のプロセスである用語の選定や構造の決定に際しては，既存の体系を参照して，必要な見出し語に漏れが無いか抽出作業の中で確認を行ったり，類似の主題部分における階層構造などを眺めてその体系に沿った観点に抜けが無いか確認することで，実装や評価が容易になり，効率的な構築が可能となる。

このような作業には，ブラウジングが有効である。そこで筆者は，語と語の関係をツリー構造で表示して眺めたり，用語間の関係を次々とたどりながら用語体系の全体像を把握したりすることにより，上記で述べた効率的な構築と，再利用が可能になると提案してきた。

4.1 用語体系ブラウジングツール

ここでは，筆者が開発した，複数の用語体系をまとめてブラウジングできるツールを紹介する [15]。

例えば，図5は，複数の用語体系（EDR，NDC，ODP）に存在する見出し語「アニメーション」について，同一表記による階層関係を重ね合わせて表示させた例である。ここで，見出し語「アニメーション」は，

図5 複数用語体系の重ね合わせ結果:「アニメーション」

- EDR における「フィルムアニメーション」の上位語
- NDC における各種映画の下位語
- ODP における「アート」の下位語
- ODP における「3D グラフィックス・ソフトウェア」の下位語

図6 用語体系 Web サービス・概念図

という、複数の体系における階層構造上の見出し語を重ね合わせたものである。このように複数の用語体系の情報をまとめて表示することで、その見出し語が持つ関係をより広い観点から一望できる。これが既存の用語体系を再利用を意図した構築プロセスの一例である。

4.2 Web Services に基づく用語体系データ提供システム

また、複数の用語体系に基づく再利用を活発に行うためには、前節の ODP の例(図3)でも示したように、用語体系データをそのまま他所で利用可能であることが必要になってくる。このようなやり方の一例として、筆者が開発した Web Services に基

づく用語体系データ提供システム [16] を紹介する。これは、それぞれの用途と目的に沿って構築された複数の用語体系を、各所で分散的に提供し、それらを統一的なインターフェースでつなぐことで、相互に連携可能な用語体系として扱えるものとなる。図6にその概念図を示す。ここでは、インターネット上の各所に分散的に独立して提供されている用語体系 Web サービス(用語体系 WS)を様々な用途に好きなように検索、加工、表示するようなフロントエンド(利用システム; 一種のアプリケーション)が自由に構築できるというものであ

図7 Google Web API との連携システム

る。用語体系はそもそも様々な組織や研究者などのニーズにより、別個に独立して作成されているという性質を踏まえ、一つずつの用語体系が、インターネット上で分散的にそれぞれが一つのサービスとして完結し、それらの間を統一的に扱えるプロトコルを定めるのが自然であると考えた。そのようにインターネット上で分散的に作成され提供されるいくつかの用語体系 WS をまとめて取扱い、目的に応じたアプリケーションとして活用する手法が望ましいと考える。想定されるアプリケーションとしては様々な形態が考えられるが、文献 [16] では試みに、各種体系をブラウジングするツールと、各種体系とサーチエンジンを組み合わせて検索の支援に利用するツールの

2種類のフロントエンドを開発した。以下、
後者のアプリケーションについて簡単に紹
介する。

図 7 はこのアプリケーションの利用画面例である。これは、用語体系データ提供システムと Google Web API [17] による検索エンジンとを連携させて、用語体系に基づく検索語選択支援機能を持たせている。例えばこのフロントエンドに検索語を渡すと、Google のサーバおよび、複数の用語体系を提供しているサーバそれぞれに、XML に基づく Web サービスのメッセージを送り、検索エンジンの検索結果を表示するとともに、用語体系の中に検索語の上位語・下位語などがあれば、それを同時に表示できる。ここでは、検索語「エネルギー」

ギー」と入れて検索を行ったところであり、サーチエンジン Google からは上位 10 件の検索結果が返ってきて表示され、同時に EDR, ODP の 2 つの用語体系から、それぞれの体系にある見出し語「エネルギー」の上位語と下位語を表示しているところである。利用者は表示された検索語の上位語、下位語を見て、検索に使いたい語があれば、その語をクリックするだけで検索欄にその語が挿入され、絞り込み検索を行うことができる。例えば、このシステムを使うと、検索者が適切な検索語が分からない場合でも、様々な用語体系が持っている語彙の中から適切なものを眺め、選択することができるので便利である。

このように、用語体系データをそのまま提供して再利用と連携を可能とすることにより、新たな用語体系の構築の際に効率的な再利用が可能となるとともに、構築された用語体系のデータの統一的なインタフェースを利用してこれまでには無い新しい形のサービスを提供することができるようになる。

5 おわりに

用語体系を巡る環境はここ 10 年で大きく変化を遂げている。指摘するまでもなく、これは主に Web を始めとするネットワークアプリケーションの環境の進展で大量の文書が蓄積されるようになったことと、コンピュータ資源がより安価で高速に

なったこととによるが、より組織化された適切な用語体系の迅速な提供が求められていると言える。とりわけ、用語体系のデータを XML などの機械可読かつその意味を伝達可能な形式で表現することにより、Semantic Web を始めとする大量情報の知的処理といった将来的な課題にも対応していけるものとする。また、その知的基盤の一つとして、ますます用語体系は重要性を増していくことが期待されており、今後その提供が速やかに行われていくことが重要になってくる。

このためには、より活発な用語体系の作成と提供が重要になると考え、本稿では、その手法として、用語体系の再利用と連携という観点から、複数の用語体系を再利用するための手法として、複数体系のマッピングによるブラウジングや、Web サービスに基づく用語体系データの提供と連携システムといった事例を紹介した。筆者は、現在の環境における用語体系をもっと広い観点から捉え、その提供をよりスムーズにする手法に取り組んでいるが、なによりも、用語体系をより活発にやりとりすることが最も重要であると考え、このためにはまず、相互運用の考え方を考慮した用語体系が提供されていくことが重要であり、その上で、それらを基盤とした連携、再利用が進められていくことを期待している。本稿がその一助になれば幸いである。

参照文献

- [1] Fellbaum, Christiane: "WordNet: an electronic lexical database", Massachusetts, MIT Press, 1998, 423p.
- [2] 日本電子化辞書研究所: EDR 電子化辞書 2.0 版仕様説明書, 東京, 日本電子化辞書研究所, 1999, TR-006, <http://www.iiijnet.or.jp/edr/>.
- [3] 池原悟他編: 日本語語彙大系, 東京, 岩波書店, 1997, 5 冊.
- [4] JICST 科学技術シソーラス 1999 年版 (CD-ROM), 1999.
- [5] Medical Subject Headings, <http://www.nlm.nih.gov/mesh/meshhome.html> (参照 2005-10-01) .
- [6] Yahoo!, <http://www.yahoo.com/> (参照 2005-10-01) .
- [7] "ODP: Open Directory Project", <http://dmoz.org/> (参照 2005-10-01) .
- [8] 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.
- [9] 日本図書館協会 件名標目委員会編: 「基本件名標目表 (BSH)」, 第 4 版, 東京, 日本図書館協会, 874p, 1999.
- [10] EIC ネット, <http://www.eic.or.jp/> (参照 2002-12-04) .
- [11] 建築工事標準分類, <http://dbnet.watanabe.arch.waseda.ac.jp/code.html> (参照 2005-10-01) .
- [12] オンライン学術用語集, <http://sciterm.nii.ac.jp/> (参照 2005-10-01) .
- [13] Alan Gilchrist: "Thesauri, taxonomies and ontologies: an etymological note", *Journal of Documentation*, 2003, Vol. 59, No. 1, pp. 7-18.
- [14] Emma McCulloch: "Thesauri: practical guidance for construction", *Library Review*, Vol. 54, No. 7, 2005, pp. 403-409.
- [15] 高久雅生; 江草由佳; 伊東千夏; 石塚英弘: データの再利用が可能な複数用語体系ブラウジングシステムの構築, *情報知識学会誌*, Vol. 13, No. 1, 2003, pp. 10-22.
- [16] 高久雅生; 江草由佳; 石塚英弘: Web サービスによる用語体系データの提供とその応用システム, *情報知識学会誌*, Vol. 14, No. 1, 2004, pp. 11-22.
- [17] Google Web APIs, <http://www.google.com/apis/> (参照 2005-10-01) .