解説

データベース支援「辞書」の紹介（同義語辞書・シソーラス辞書）

Introduction of Dictionary for Database System Aiding (Dictionary of Synonym・Thesaurus)

河野　弘
日本データベース開発㈱

【背景】

用語の標準化、辞書の必要性は、テキスト系データベースの出現以来早くからその必要性が認識されてきたが、当初は索引語の統一によるデータ検索の精度アップというところにその主目的があり、一部専門的グループによってのみ、必要性・目的に応じて用語の体系管理が行なわれてきた。（概念の階層を反映したシソーラス辞書も一部機関にて作成された実績があり、データへの索引語付与の高精度化とDB検索の柔軟性の確立に貢献するものであった）

しかしながら、用語の体系的整理は時間と費用を膨大に必要とするのが常であり、規模の大きさを伴う標準的・体系的辞書の作成は見送られるケースが多いまま、今日に至っている。

一方、情報インフラの爆発的拡大、データベース構築用簡易ツールの出現、高機能検索エンジンの一般化という流れの中で、DB用「辞書」の問題は今日新たな課題として、必要性に応えるべき状況が発生していると言える。

①インターネット系を含め、DB検索機会の増加（データ検索の効率化）…図書館でのOPAC検索や社内DBの活用
②用語の累乗的増加（専門語、外来語、新語など）へのシステム側の対応
③バイリンガル環境の進展（情報のグローバル化）
④テキストマイニング技術などによる辞書の必要性

等がその背景事情にあると言える。

こうした背景の中で、最近に至りDB支援のための辞書作りが一部始められている。

汎用機（コンピュータ）のもとに、高度な辞書作成専用システムを作り、人手を中心とした辞書作りの時代には数十万語という用語を、体系化することは費用・労力とも大変であった。

PCの発展と一般化に伴い、現在では簡易的なツール群（用語の切り出し、整理、対訳付与などの補助ツール）が出現し、目的と分野を明確にされた条件下での辞書は、比較的簡単に作れる状況になっている。

これより紹介する「辞書シリーズ」は、DB検索の補助となるべき「DB支援辞書」であるが、具体的用途として「ヒット率」向上のためのみでなく、検索の条件選択、ノイズの排除、バイリンガルでの検索、関連用語の通覧などを可能にし、DB関連テクノロジーの新たな用途を拓く可能性を秘めていると言える。

辞書その１.

名称【技術同義語辞書】

１．特徴

専門技術用語の整理を行った用語体系ファイルであり、55万語の標準技術用語を分野別に整理し、該当用語に検索同義語を持たせている。（用語、同義語のそれぞれが対訳英語を有する。）

2．分野構成・語数

分野名	用語数	備考(主な該当分野)
技術一般	105,000	分野を限定しない技術用語、略語、表記ゆれ語
地球	33,000	海洋、地球、地質、地理、気候、宇宙
環境	11,000	汚染、廃棄物、騒音、動植物被害、環境技術
物理	41,000	音響、振動、流体、光、磁気、素粒子、放射線
エネルギー	9,000	原子力、自然エネルギー、エネルギー利用技術
エレクトロニクス	63,000	情報、コンピュータ、電気、通信、電子
機械	61,000	産業機械、精密機械、車両船舶航空機
ケミストリー	82,000	無機、有機、高分子、油脂、繊維、写真、印刷
医薬・バイオ	116,000	医学・薬学、生物、遺伝、動植物、農林水産、食品
マテリアル	23,000	金属、非鉄金属、加工技術、合金、鉱山技術
建設	11,000	建設土木、土質、材料、都市計画、防災、交通
(合計)	555,000	

3．同義語の内容

異表記同義語、略語による同義語、意味的同義語、検索類義語をサポート(同義語保有率は全体用語の50%程度)

用語・同義語・対訳英語は後述「用語サンプル」を参照

4．辞書の構造と提供形態

提供形態:CSV ファイル

(用語, 対訳英語, 同義語①, 対訳英語①, 同義語②, 対訳英語② ……)

辞書構造:以下の Excel フォーマットにて管理が行なわれている。(分野別)

用語番号	用語	対訳	同義語①	対訳②	同義語②	対訳②	同義語③	対訳③

＊同義語の保有は1用語に5語以内

5．メンテナンス

年1回(新語の追加、同義語の追加)

6．拡張性(ユーザーカスタマイズを含む)

1) 各分野内での用語階層設定(分類内「用語」への小分類設定)
2) 英語以外の対訳語付与(中国語など)
3) 用語への注釈(簡単な説明)付与
4) 検索エンジンへの取込み

7．用途目的

1) 検索ヒット率の向上(各種全文検索システムへの組込みにより、キーワードの通覧・選択機能、同義語検索支援機能などに辞書適用)
2) 英文情報への「日本語キーワード」によるアクセス(用語対訳を適用)
3) ハイパーリンクによる、DB 支援への適用(用語の対訳・同義語・注釈などをテキスト画面上にリンク表示)

4) 自動インデックスツールの対応辞書としての用途（用語の網羅性）

【用語例】1

用語(対訳英語)	同義語(対訳英語)	備考
4WD (four wheel drive)	四輪駆動 (four wheel drive) 四輪駆動車 (four wheel drive) 4輪駆動 (four wheel drive) 4輪駆動車 (four wheel drive)	
BSE (bovine spongi form encephalopathy)	狂牛病 (bovine spongi form encephalopathy) うし海綿状脳症 (bovine spongi form encephalopathy)	
トモグラフィー法 (tomography)	断層撮影 (tomography) 断層X線撮影 (tomography) 断層エックス線撮影 (tomography)	
男性ホルモン (androgen)	雄性ホルモン (androgen) 雄ホルモン物質 (androgen)	
地すべり(land slide)	地滑り (land slide) 落盤 (land slide)	

辞書その２.

名称【ヨミダス】(読売新聞シソーラス辞書)

１. 特徴

新聞記事上で使われる用語を整理し、記事検索の上で幅広いカバーが可能となるよう、「記事上の用語」＊「検索者の思いつく言葉」のずれを埋めることを主目的としている。

記事検索のための連想語辞書ということもできる。

用語のカバー分野は、政治・社会・文化一般など新聞出現用語が中心に、全カテゴリー分野をカバーするものである。

２. 分類（分野構成）

（各用語は、14の大分野とそれに付随する135の小分野にカテゴライズされている。また固有名詞は8分類に体系化されている。）…全 54,000語

1) 大分野カテゴリー【政治、経済、産業、社会、生活家庭、事件、労働、教育、文化、科学、スポーツ、皇室、国際、その他】

2) 固有名詞カテゴリー【人名、企業名、国内機関団体名、法令名、国内地名、競技名、国際機関団体名、世界の国名地名】

３. 辞書の構造と提供形態

提供形態:CSV ファイル

（用語, 同義語…, 狭義語…, 関連語…）

辞書構造

以下の Excel フォーマットで管理されている。

用語	ヨミ	狭義語	関連語	同義語	分野

同義語・狭義語・関連語内容は後述「用語サンプル」を参照

4．メンテナンス

年1回の更新（新語採択、シソーラスメンテ）

5．拡張性

1）シソーラス機能の充実
（各種表記ゆれ、略語、など）

2）文字コードの対応拡張（現在はシフトJIS）

6．用途目的

1）新聞記事DB検索への組込み（読売新聞記事CD版に適用済み）
2）一般DB検索、図書検索などへの適用

【用語例】

用語	狭義語	関連語	同義語
警察官	巡査、巡査長、羅卒、巡査部長、代警視、警部、警部長、中警視、警部補、警視補、小警部、権中警視	警視庁、警視局、探偵駐在、刑事、ニセ刑事	警官
書籍	絵本、参考書、辞典、児童書、新刊書、新刊図鑑、単行本、専門書年鑑、文庫本、美術書社史、有害図書、育児書、ビジネス書		図書 書物 本
建国記念日	紀元節	神武天皇	
赤穂浪士	赤穂義士	泉岳寺、忠臣蔵	
赤ん坊	赤子、赤児、乳児、みどりご、緑児	嬰児、えい児	
狂牛病	異常プリオン、変形型クロイツフェルト・ヤコブ病	肉骨紛、配合飼料、牛、水牛、神経症状	牛海綿状脳症 BSE