情報検索の語彙を拡張するためのシソーラスの統合

大村玲子（慶應義塾大学大学院）reiko.omura@a5.keio.jp

１．はじめに

情報検索の手段には、自然言語と統制語彙がある。シソーラスに代表される統制語彙は、自然言語検索における語彙や構文上の短所を補い、検索性能を高める目的で、1950 年代以降の情報検索の歴史に伴い変遷を辿っている。

特に 1960～80 年代のオンライン検索時代にはデータベース単位のシソーラスが急増し、1990 年代以降のインターネット時代には急成長した分散レポジトリに対し新たな形の統制語彙が台頭した。結果、情報の横断的検索のためには、シソーラスの統合が急務となり、研究・運用が進んだ。シソーラス構築の標準化はこの点でも重要な役割を担い、従来の単一言語を主に扱う国際／国家標準が、「相互運用性」へのニーズに応える形で近年見直されている。

本研究は、第 1 に、シソーラスの統合に注目し、その特徴や各種方法論を整理展開するものである。同時に、近年見直されたシソーラスの標準規格の内容のうち、統合に関連する事項を議論する。第 2 に、事例として、スポーツ分野の海外領域シソーラスと国内の標準件名標目表の統合を試み、その課題を探究する。

２．シソーラスと統合

２．１　統制語彙の種類

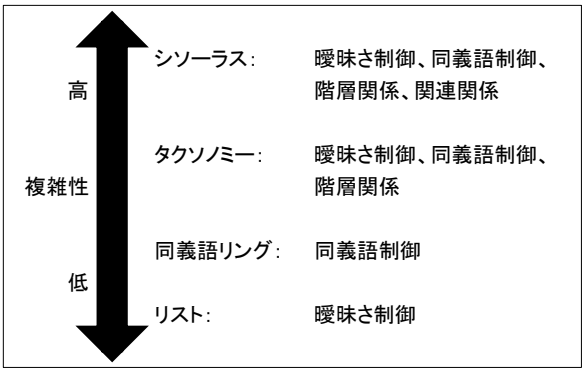
国際標準化機構（ISO）は、7 種の統制語彙と概念の表現要素を表 1 にまとめている [1)]。

表 1. 統制語彙の種類と概念の表現要素

種類	概念の表現要素
シソーラス	優先語
分類表	記号
タクソノミー	分類ラベル／記号
件名標目表	見出し
名称典拠リスト	名称
オントロジー	ラベル
用語集	用語／他種の記号

また、米国規格（ANSI/NISO Z39.19）では、構造の複雑性の観点から、4 種の統制語彙を図 1 で表現している [2)]。

図 1. 統制語彙の種類と構造の複雑性



２．２　統合の種類

シソーラス、分類表などの統制語彙の統合、併合、調整やマッピングに関する議論の歴史は 1970 年代に遡る [3)]。その歴史を時系列に概観すると、Aitchison & Gilchrist（1989）は、特定性、網羅性、複合語、同義語、相互関係（階層、関連関係）の相違による互換性問題を調整する方法として、①変換システム、②併合と統合、③マクロシソーラスとマイクロシソーラスを提示している。また、彼らが多言語シソーラスを詳細に取り上げている点は見逃せない [4)]。

次に、Zeng と Chan は 2000 年以降、統制語彙をインターネット時代に即した「知識組織化体系」（Knowledge Organization System: KOS）と捉え、KOS 間の相互運用性確保のための統合と方法論を報告している。2002 年時点では、①派生／モデル化、②翻訳／翻案、③マッピング（人手）、④マッピング（コンピューター支援）、⑤リンキング、⑥スイッチングの 6 種に統合方法を分類した [5)]。その 2 年後には、①派生／モデル化、②翻訳／翻案、③サテライト・リーフノードリンク、④ダイレクトマッピング、⑤共起マッピング、⑥スイッチン

グ、⑦仮のユニオンリストを利用するリンク、⑧シソーラスのサーバプロトコルを利用するリンク（下線が変更点）の 8 種に分類を詳細化及び追加している[6]。因みに、2004 年版の統合モデルは、図を含め ANSI/NISO Z39.19 規格の付録に参照されている[2]。

更に、標準規格 ISO 25964-2 は、①構造の統一、②ハブ構造、③選択マッピング、④ダイレクトリンク（ダイレクトマッピング）の 4 つの統合の種類を紹介している[1]。

２．３　Zeng & Chan の統合モデル

本研究では、KOS の同種・異種間の統合に言語を介在させた事例を多く分析した 2004 年の Zeng & Chan の統合モデルを基本とする。

中でも、事例（後述）で扱う以下の 4 種のモデルにつき、簡単に説明する。

（１）翻訳／翻案

既存の統制語彙を翻訳或いは翻案し、新たな統制語彙を構築する。その際、基本的に全体的な構造や指針は既存のものに準じる。

例：RVM(ｶﾅﾀﾞ)始め多くの LCSH の翻訳など

（２）サテライト

スーパーストラクチャー(上層)構造に対し、サテライトの関係を持つ統制語彙を構築する。結果、専門性に富んだ統制語彙が期待できる。

例：GLIN（米）、Getty Project（米）など

（３）ダイレクトマッピング

異なる統制語彙間の用語の等価関係などを対応づけて統制語彙を構築する。通常、人手による高度な知識と技術が必要とされる。

例：LCSH/MeSH（米）、Marimee（仏）など

（４）スイッチング

ある言語や体系を交換手段として介在させ、異なる統制語彙間の用語の対応づけを行う。交換手段は既存のものでも新たなものでもよい。

例：H.W. Wilson (米)のメガシソーラスなど

シソーラスの統合における主な課題は、シソーラス自体に潜在する複雑性、統制語彙の種類の増加と異種間の互換性の低さ、そしてグローバル化に伴う多言語化要求に起因する。果して、情報技術と統合モデルの発展は、利用者の検索を容易にし、検索性能を高めるであろうか。

３．シソーラスの標準化

シソーラスに関わる標準規格（欧州除く）には以下のものがある。

（１）国際標準規格

ISO-25964-1: 2011 (Part 1)

ISO-25964-2: 2013 (Part 2)

（２）国家標準規格

ANSI/NISO Z39.19: 2005 (2010 改訂)（米国）

JISX0901: 1991（日本）

（３）その他

IFLA Guidelines for multilingual thesauri (2009)[7]

国際／国家標準である ISO、ANSI/NISO とも、近年シソーラスの構築に関する規格内容を見直した。従来参照されていた英国規格の BS 8723（2005）は ISO-25964-1 に置き換わり、加えて ISO-25964-2 がシソーラスと他の統制語彙との相互運用性に特化して分冊出版された。また、ANSI/NISO も同様に 2005 年に改版し、2010 年に改訂を行った。ISO に比べると、相互運用性を記述している量は少ないが、両者とも、電子シソーラス・多様なユーザー(エンドユーザー、索引作成者、専門家、開発者など）の利用を前提に、時代に見合った統制語彙を拡張し（表 1、図 1 参照)、更にユーザーインターフェースや表示方法の改良点を盛り込んでいる。

４．事　例

４．１　利用する統制語彙

本研究の事例として、スポーツ分野の領域シソーラス、*SIRCThesaurus 6* (2002)[8]（カナダの Sport Information Resource Center, SIRC

が 2005 年まで SPORTDiscus データベースに用いていたシソーラス）と、『基本件名標目表（BSH）第 4 版』[9]を利用する。国内でのスポーツ分野のシソーラスの開発は、1980 年代末に鹿屋体育大学で研究・報告されている[10]が、現在公式に利用されている日本語のシソーラスは存在しない。

４．２　目的と方法

上述のスポーツ領域シソーラス（英語）と件名標目表（日本語）を統合し、日本語の領域シソーラスを構築し、最終的にはエンドユーザーがスポーツ分野の文献を検索する際の支援となるような検索シソーラスを作ることを目的とする。

ここでは実際に、特定領域と全領域という異なる領域間、シソーラスと件名標目表という異なる統制語彙間、そして日本語と英語という異なる言語間での統合を試みることになる。これは、２．３節で述べた統合モデル、①翻訳／翻案（英語から日本語）、②サテライト（件名標目表よりスポーツ分野のみ適用）、③ダイレクトマッピング（シソーラスと件名標目表双方向の統合）、そして④スイッチング（言語を介し、NDC を参照する）を採用することを意味する。

４．３　統合の経緯と課題

シソーラスと件名標目表とは、シソーラスが特定の主題分野を扱い、ディスクリプタで主題の構成概念を表すのに対して、件名標目表は全主題分野を網羅的に扱い、件名が主題全体を表すという点で異なっている。また、前者は事前結合のレベルが低く、階層構造や関連関係は十分に設定されるが、後者は事前結合のレベルが高く、階層構造や関連関係は十分に設定されない[11]。

例にもれず、SIRCThesaurus（以下、SIRC）と BSH の関係も、これらの特徴ゆえに、語彙の拡張においては統合が有効に働くが、階層の相違に加え、翻訳の介在が統合を困難に導く。

（１）階層と語彙の拡張（成功例）

例えば、ディスクリプタ「スポーツ」に対し、SIRC は NT（下位語）18 語のうち、16 語（89%）が階層構造を持つ（更に NT や RT を持つ）のに対し、BSH は NT45 語のうち階層構造を持つ語は 13（29%）のみである（表 2）。「スポーツ」の例に見られるように、スポーツ分野の語彙数は圧倒的に SIRC が勝り（約 27,000 語）、従って両者の統合は BSH（全分野で 8000 件弱）の弱点を補い、語彙を拡張する結果になる。

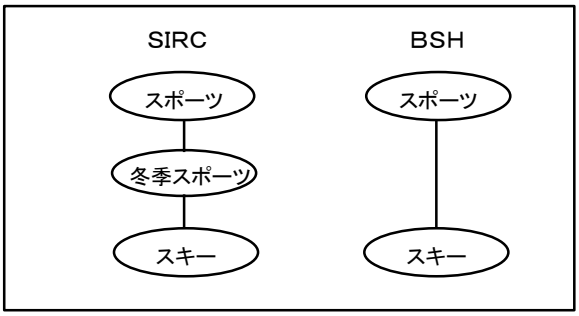
表 2. SIRC(左)と BSH(右)：「スポーツ」の比較

スポーツ（SIRC）		スポーツ（BSH）	780（NDC）
UF	身体活動	UF	運動競技
	競技スポーツ		競技
	視覚障害者スポーツ	NT	アジリティードッグ 788.4
	大学対抗競技		オリエンテーリング 796.5
	学校対抗競技		■オリンピック 780.69
	サイレントスポーツ		カバディ 780.7, 788
	スポーツ管理		カーリング 794.9
	スポーツマネジメント		■球技 783
	スポーツ博物館		競技場 780.7
NT	■航空スポーツ		コーチ（スポーツ） 780.7
	■アニマルスポーツ		サイクリング 796.5
	■水上競技		■射撃 789.7
	■球技、特に野球		重量挙げ 782.8
	■格闘技		審判（スポーツ） 780.7
	■サイクリング		■水上競技 785
	■エクストリームスポーツ		スキー 794.3
	■体操		スキューバダイビング 785.3
	■生涯スポーツ		■スケート 794.6
	ミニスポーツ		スケートボード 796.8
	■モータースポーツ		スノーボード 794.4
	スポーツツーリズム		スノーモービル 794.9
	■ターゲットスポーツ		■スポーツ医学 780.2
	■団体競技		スポーツ事故 780
	■陸上競技		■スポーツ心理学 780.14
	■重量挙げ		スポーツ選手 780
	■車椅子スポーツ		スポーツ団体 780.6
	■冬季スポーツ		スポーツルール 780
RT	遊戯		■相撲 788.1
			潜水術 785.3
	■：　階層あり		■体操 781
	●：　同じ語彙あり		チアガール 781.8
			綱引き 782
			■登山 796.1
			トライアスロン 782.6
			バイアスロン 794.4
			■馬術 789.6
			パラグライダー 782.9
			ハンググライダー 782.9
			■武道 789
			フライングディスク 796.9
			ボクシング 788.3
			ボディ　ビル 781.5
			マスゲーム 781.8
			ランニング 782
			■陸上競技 782
			レスリング 788.2
			ローラースケート 796.8
		RT	■体育 780

（２）階層構造の相違（問題例）

階層の深さと語彙の多さが特徴の SIRC であるが、BSH にない「スポーツ」の下位語に注目すると、2 つの特徴がある。1 つは、「ターゲットスポーツ」「団体競技」「冬季スポーツ」など、ファセット機能を持つ語彙が散在することである。単純な例として、ディスクリプタ「スキー」は、図 2 のとおり階層構造に違いがある。

図 2. 「スキー」の階層比較



2 つ目に、文化社会的背景に起因する語彙が多く、階層にも影響することである。例えば、「アニマルスポーツ」「スポーツツーリズム」「車椅子スポーツ」などである。スポーツは国際的活動ではあるが、文化社会的影響を受けやすい分野でもある。日本語のスポーツシソーラスを構築する際に、ノイズとなりかねないこれらの語彙や階層をどのように扱うかは課題となる。

（3）翻　訳（問題例）

スポーツ分野の語彙はスポーツの種目を表すものが多く、必然とカタカナ表記による翻字が多い。和訳・字訳する際には、単純な翻訳作業とは別に、BSH や時に NDC の語彙と照合するなど文脈を汲む翻訳作業が必要となる。領域用語の英和・和英辞書が十分揃っていない本研究の事例などにはとりわけ、多言語シソーラスの統合における翻訳の難しさが存在する。

５．おわりに

インターネット時代の情報検索において、シソーラス機能は新たな意義と役割を期待されている。本研究では、Zeng & Chan の統合モデルを数種適用し、スポーツ分野に特化したシソーラスの統合を試みた。その結果、階層構造の相違と翻訳に関わる統合上の問題が明らかになった。今後はこれらの問題を更に洗い出すことにより原因を追究し、シソーラスを統合する手段を確立することを本研究の課題とする。

引用文献

1) ISO 25964-2 (2013) : Information and documentation : thesauri and interoperability with other vocabularies : Part 2, Interoperability with other vocabularies. ISO.

2) “ANSI/NISO Z39.19-2005 (R2010) : Guidelines for the construction, format, and management of monolingual controlled vocabularies”. NISO. (accessed 2014-08-24) http://www.niso.org/apps/group_public/project/details.php?project_id=46

3) Shiri, Ali. Powering search : the role of thesauri in new information environments. Information Today, 2012, 318p., (ASIST monograph series)

4) Aitchison, Jean; Gilchrist, Alan. シソーラス構築法. 第 2 版, 内藤衛亮ほか訳. 丸善, 1989, p117-127.

5) Chan, Lois Mai; Zeng, Marcia Lei. Ensuring interoperability among subject vocabularies and knowledge organization schemes : a methodological analysis. IFLA Journal. 2002, no.28, p.323-327.

6) Zeng, Marcia Lei; Chan, Lois Mai. Trends and issues in establishing interoperability among knowledge organization systems. Journal of the American Society for Information Science and Technology. 2004, vol.55, no.5, p.377-397.

7) “IFLA Guidelines for multilingual thesauri”. IFLA, http://www.ifla.org/publications/ifla-professional-reports-115 (accessed 2014-08-24).

8) SIRCThesaurus 6. Sport Information Resource Centre, 2002.

9) 基本件名標目表 (BSH). 日本図書館協会, 1999, 第 4 版.

10) 池田勝ほか編. 体力・スポーツ科学に関するデータベースと文献情報検索システムの開発に関する研究. 鹿屋体育大学, 1990, 119p., (昭和 63 年度文部省科学研究費研究成果報告書).

11) 岸田和明. 情報検索の理論と技術. 勁草書房, 1998, p.46-47, (図書館・情報学シリーズ 3).