人文学分野への大規模用語辞書の応用

甲田 彰
独立行政法人 科学技術振興機構 情報提供部

梶 正憲
独立行政法人 科学技術振興機構 文献情報部

森田 歌子
独立行政法人 科学技術振興機構 審議役

科学技術振興機構（以下「JST」）は2006年4月から，新たな科学技術文献検索サービス「JDreamⅡ」の提供を開始した．JDreamⅡには様々な検索支援機能を搭載しており，そのうちの一つが今回報告する大規模用語辞書およびそれを用いたJSTシソーラスブラウザである．本稿では，JDreamⅡの概要と大規模用語辞書の活用方法を述べたうえで，人文学分野への展開の可能性を探る．

Application of a large-scale scientific and technological dictionary to the humanities study

Akira Kouda
Department of Service, Japan Science and Technology Agency(JST)

Masanori Kaji
Department of Literature Information, Japan Science and Technology Agency(JST)

Utako Morita
Executive Director, Japan Science and Technology Agency(JST)

We, Japan Science and Technology Agency(JST) has released a new science and technology document retrieval system named JDream II for JST database on April 2006. For wide users, JST supports several new functions, for example the JST thesaurus browser based on a large-scale scientific and technological dictionary. This manuscript describes the usage of JDream II and the JST thesaurus browser, and possibility of application to the humanity study.

1．まえがき

科学技術振興機構（以下「JST」）は2006年4月から，新たな科学技術文献検索サービス「JDreamⅡ」の提供を開始した．JDreamⅡには様々な検索支援機能を搭載しており，そのうちの一つが今回報告する大規模用語辞書およびそれを用いたJSTシソーラスブラウザである．

大規模用語辞書やJSTシソーラスブラウザは，科学技術文献データベースでの利用を目的に作成した仕組みではあるが，そのコンテンツや仕組みは，科学技術文献に限らず利用可能と考える．本稿では，JDreamⅡの概要，大規模用語辞書の活用方法を報告したうえで，人文学分野への展開の可能性を探る．

2．JDreamⅡの概要

2．1．JDreamⅡのデータベース

JSTは，国内外の科学技術文献と国内の医学関連文献を収集し，書誌データや概要（抄録）の作成，検索用索引語の付与等の加工を行い，文献データベースとしてJDreamⅡで提供している（表1）．また，他の機関で作成した情報もデータベースとして提供している(表2)．

データベース名	収録情報	収録年代（更新頻度）	収録件数
JSTPlus	科学技術（医学を含む）全分野に関する文献情報．世界５０数カ国の情報を含む．	1981－（月４回）	約1,869万件
JST7580	科学技術全分野に関する文献情報．世界５０数カ国の情報を含む．	1975－1980（更新無し）	約220万件
JMEDplus	日本国内発行の資料から医学，薬学，歯科学，看護学，生物科学，獣医学等に関する文献情報を収録．	1981－（月２回）	約393万件
JCHEM	化学物質の商品名，治験番号，体系名，化合物辞書番号，CAS登録番号，分子式などの情報．	（月１回）	約236万件

（表１）JST作成データベース

データベース名	収録情報	収録年代（更新頻度）	収録件数
MEDLINE	米国国立医学図書館（NLM：National Library of Medicine）が作成・提供する医学およびその関連領域を対象とする文献情報.	1966－（週１回）	約 1,442 万件
JAPICDOC	日本医薬情報センターが作成・提供する医薬品の有効性，安全性に関する文献情報	1983-（月１回）	約 31 万件

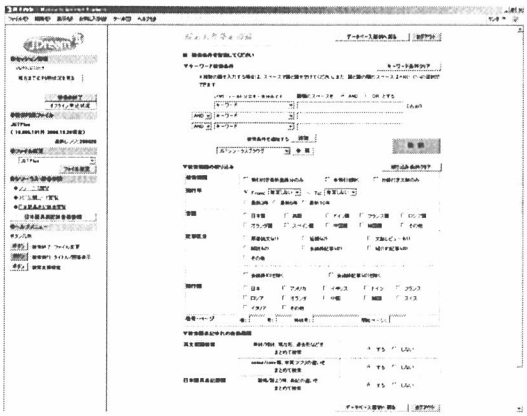
（表２）他機関のデータ

2. 2. JDreamⅡの概要

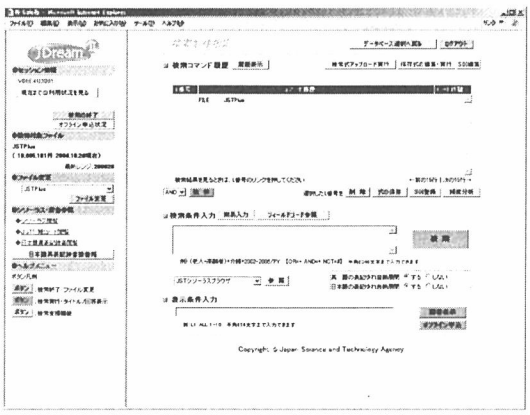
JDreamⅡは上記の文献データベースを検索するための WEB ベースのシステムであり，我が国の科学技術文献検索サービスで最も利用されているシステムである.

JDreamⅡには，エンドユーザ向けの「シンプルモード」と検索の専門家向けの「コマンドモード」を用意した.

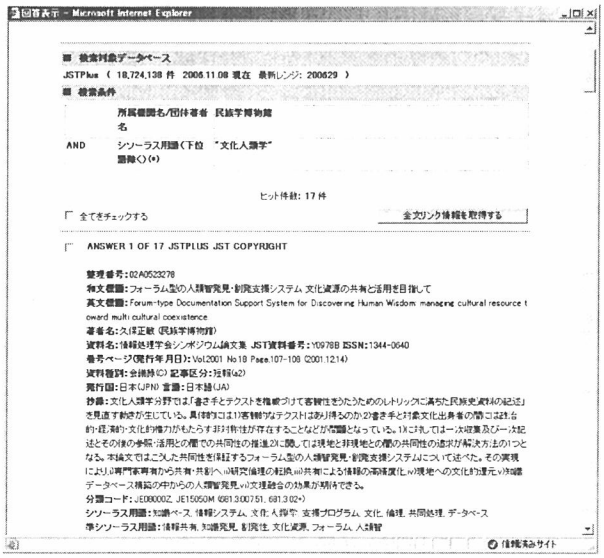
シンプルモードは検索の結果に対して検索条件の追加，頻度分析結果による絞り込み検索が可能で，初心者でも迷わず使用できる「直線的」なインタフェースである（図１）．一方，コマンドモードは，検索結果の集合間の論理演算や検索式の保存・再利用が可能な「試行錯誤型」のインタフェースである（図２）．いずれのモードでも科学技術文献情報の表示（図３）は共通で，ユーザの習熟度や利用状況に応じた使用が可能である.



（図１）シンプルモード



（図２）コマンドモード



（図３）データの表示

3. JDreamⅡにおける大規模用語辞書の活用

3. 1. 大規模用語辞書利用の目的

JST が提供するデータベースは，索引語を用いて検索することにより，漏れのない精度の高い検索を行うことが可能である．しかしながら，索引語を使用して検索するためには，そもそも索引語がどういった体系で構築されているか，どのような意味で使用されているか，等を予め知っておく必要がある．

JST が過去に提供していたシステムには索引語の参照機能を搭載していたが，ユーザがあらかじめ索引語に見当をつけて参照する方法を採っており， JST シソーラスを熟知したユーザしか効果的な検索を実施することはできなかった．

これに対して，JDreamⅡでは，ユーザがより簡便に索引語を見つけだすことができるよう，科学技術関連用語の同義語や異表記語を収録した大規模用語辞書と，それを JDreamⅡのユーザが閲覧するための専用のツール「JST シソーラスブラウザ」を用意した．JST シソーラスブラウザを用いて大規模用語辞書を参照することで，JDreamⅡのユーザは漏れなく高い精度の検索が可能となった．

3. 2. 大規模用語辞書の内容

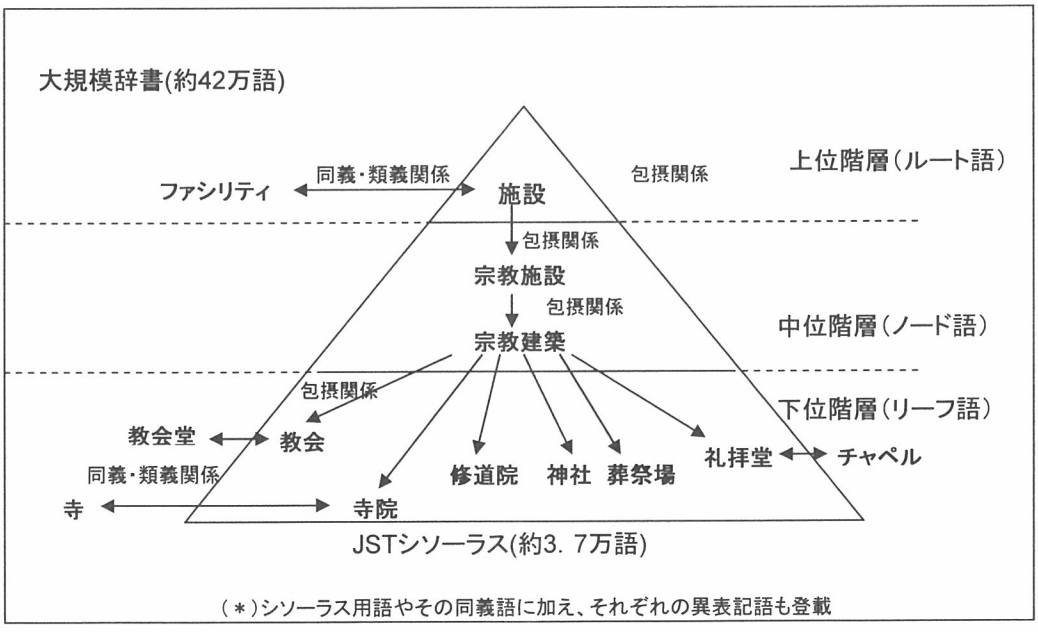
大規模用語辞書の 2006 年 11 月時点の収録用語数は約 42 万語であり，収録内容は以下の通りである（図４）．

（１）JST シソーラスや各種事典・用語集に収録されている用語とその同義語，異表記語

（２）自然語と JST シソーラスとの包摂関係も一部収録

（３）文献データベースにおける自然語の索引付けにおいて出現頻度の高い用語も収録

なお，JST シソーラスには，表２の通り人文学系のシソーラス用語も登録されている．



(図４)大規模用語辞書と JST シソーラス

カテゴリ	JST シソーラス用語（抜粋）
社会・文化（ID01）	演劇，音楽，絵画，楽譜，家族，過程，芸術，考古学，社会科学，宗教，心理学，生活，美学，風俗，文化人類学，民俗学，倫理学，…
建築設計（AB04）	駅舎，音楽堂，会議場，記念建造物，宮殿，教会，競技場，劇場，建築史，講堂，古建築，寺院，宗教建築，修道院，城，神社，葬祭場，美術館，礼拝堂，…

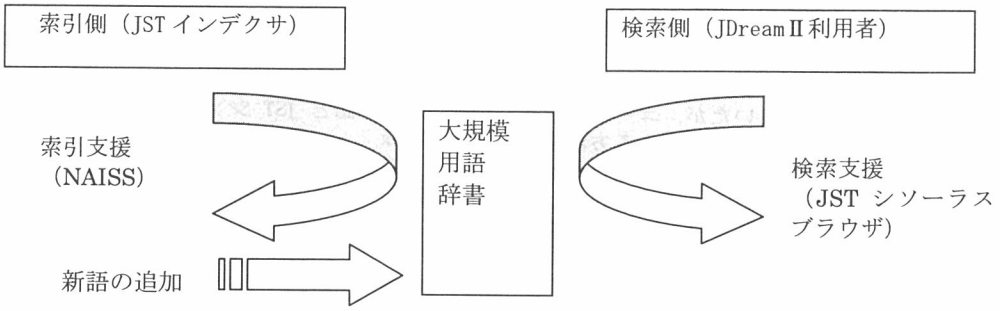
(表２)人文学系の JST シソーラス用語(抜粋)

3. 3. 大規模用語辞書の特徴

大規模用語辞書は，収録用語数の規模，種類だけではなく，索引側と検索側が同じ辞書を共有する点に特長がある（図５）．

索引側は，索引支援システム（NAISS）を用いて，抄録やタイトルから，重要と思われる単語を抽出し，大規模用語辞書を用いて索引語を付与する．大規模用語辞書にない新語を見つけた場合は辞書メンテナンスシステムを用いて大規模用語辞書に単語を追加する．

一方，検索側は，JST シソーラスブラウザを用いて，思いついた単語にふさわしい索引語を参照し，JDream Ⅱの検索で使用することが可能である[1]．



（図５）索引側と検索側からみた大規模用語辞書

4. 人文学分野への大規模用語辞書の応用

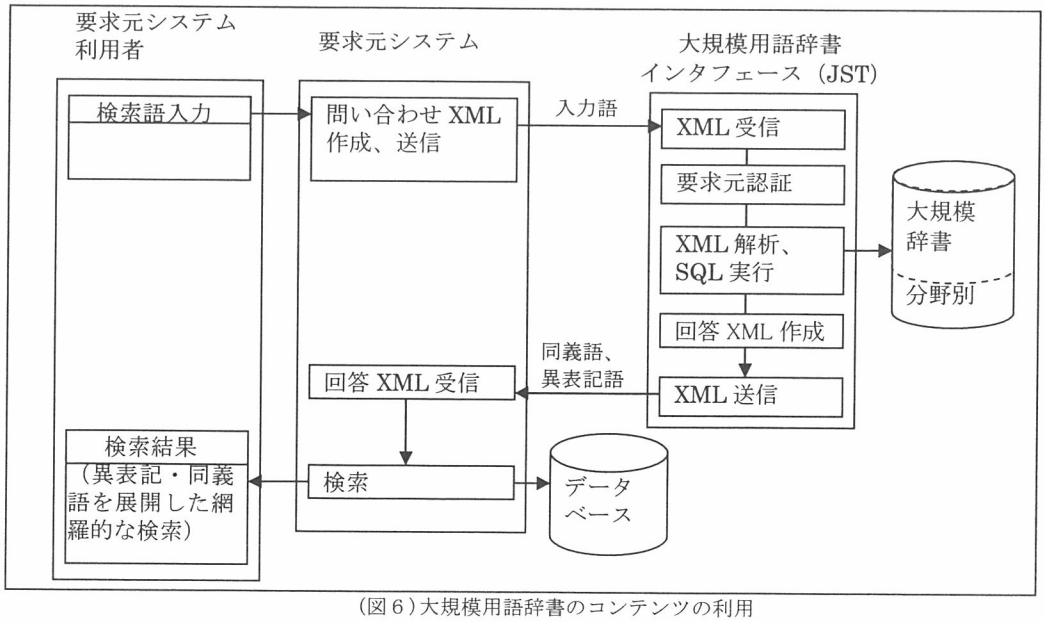
人文学分野における大規模用語辞書の展開の可能性について，２つの方向性を探る．

（１）JST の大規模用語辞書の利用

（２）大規模用語辞書構築手法とシソーラス作成への展開

4. 1. JST の大規模用語辞書の利用

大規模用語辞書に対するインタフェースは，JST の索引者が使用するシステム（NAISS）と JDream Ⅱのユーザが使用する JST シソーラスブラウザのみであるが，将来的に大規模用語辞書に対する汎用的なインタフェースを構築することも検討しており，これが完成すれば，大規模用語辞書に含まれる異表記語と同義語語を他のシステムで利用することが可能である．システムの連携を容易にするため，インタフェースを XML ベースとする等の工夫も考えられる（図６）．

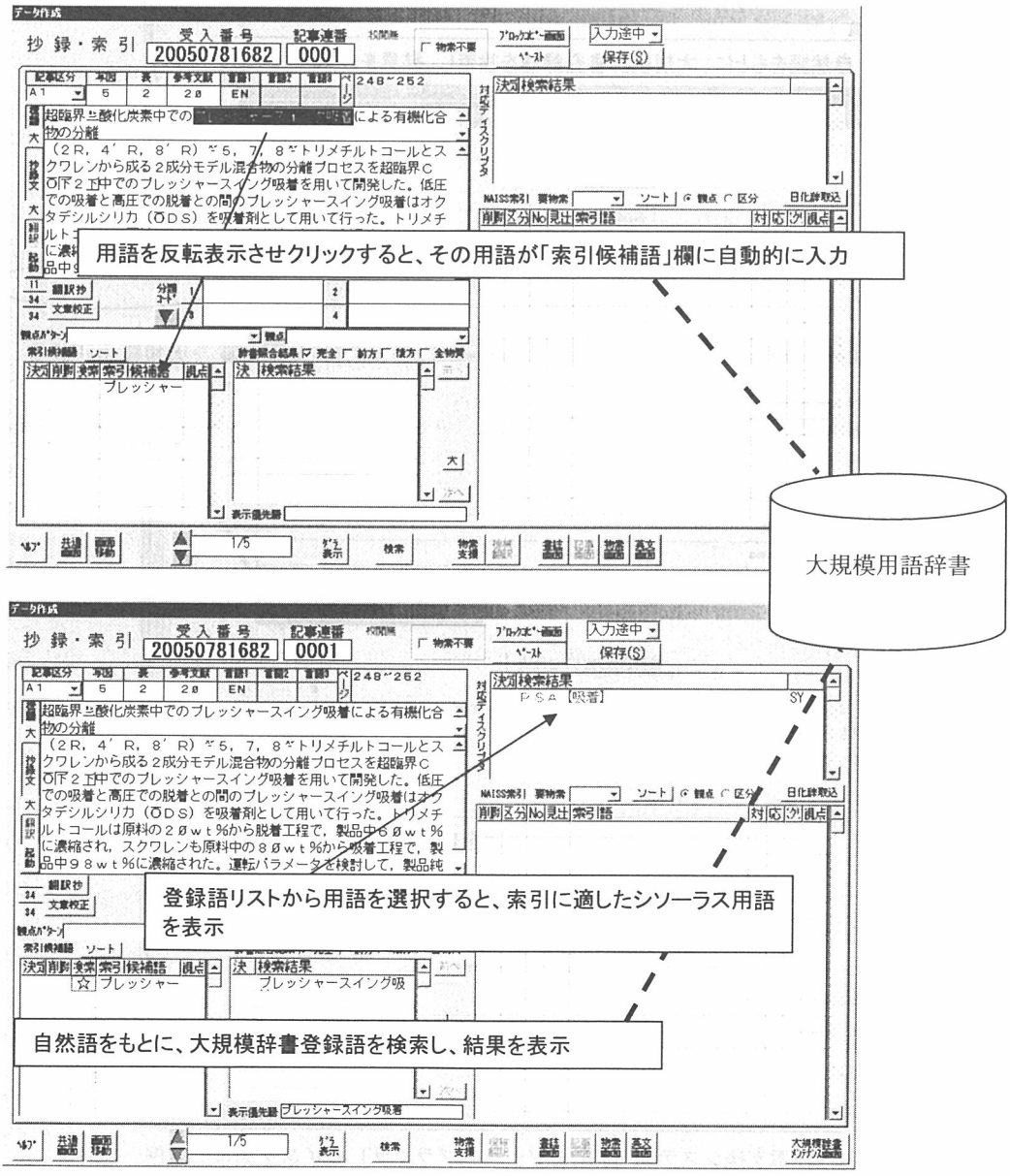


（図６）大規模用語辞書のコンテンツの利用

4. 2. 大規模用語辞書構築手法とシソーラス作成への展開

4. 2. 1. 索引支援システム，検索支援システムの検討材料

すでにシソーラス用語やその異表記語・同義語情報を持っていて，今後索引支援システムや検索支援システムでの活用を予定している場合，大規模用語辞書を元に構築した JST の索引支援システム，検索支援システムを検討材料として利用することが可能と考えられる．JST 場合，いずれも実際の索引作業者，検索ユーザからの様々な要望をもとに作成したもので，基本的なインタフェース，画面遷移等は分野によらず活用可能と考える．図７に索引支援システム（NAISS）のインタフェース，図８に検索支援システム（JST シソーラスブラウザ）のインタフェースを記載する．



（図７）索引支援システム（NAISS）のインタフェース例

（図８）検索支援システム（JST シソーラスブラウザ）のインタフェース例

4. 2. 2. シソーラス構築補助

シソーラスの構築には様々な手法があるが，大規模用語辞書を一旦作成し，それを元にシソーラスを構築することが可能である．

現在，JST では大規模用語辞書に基づくシソーラスの改訂を進めている．以下に大規模用語辞書の構築方法とシソーラス改訂作業における利用方法を説明する（図９）．

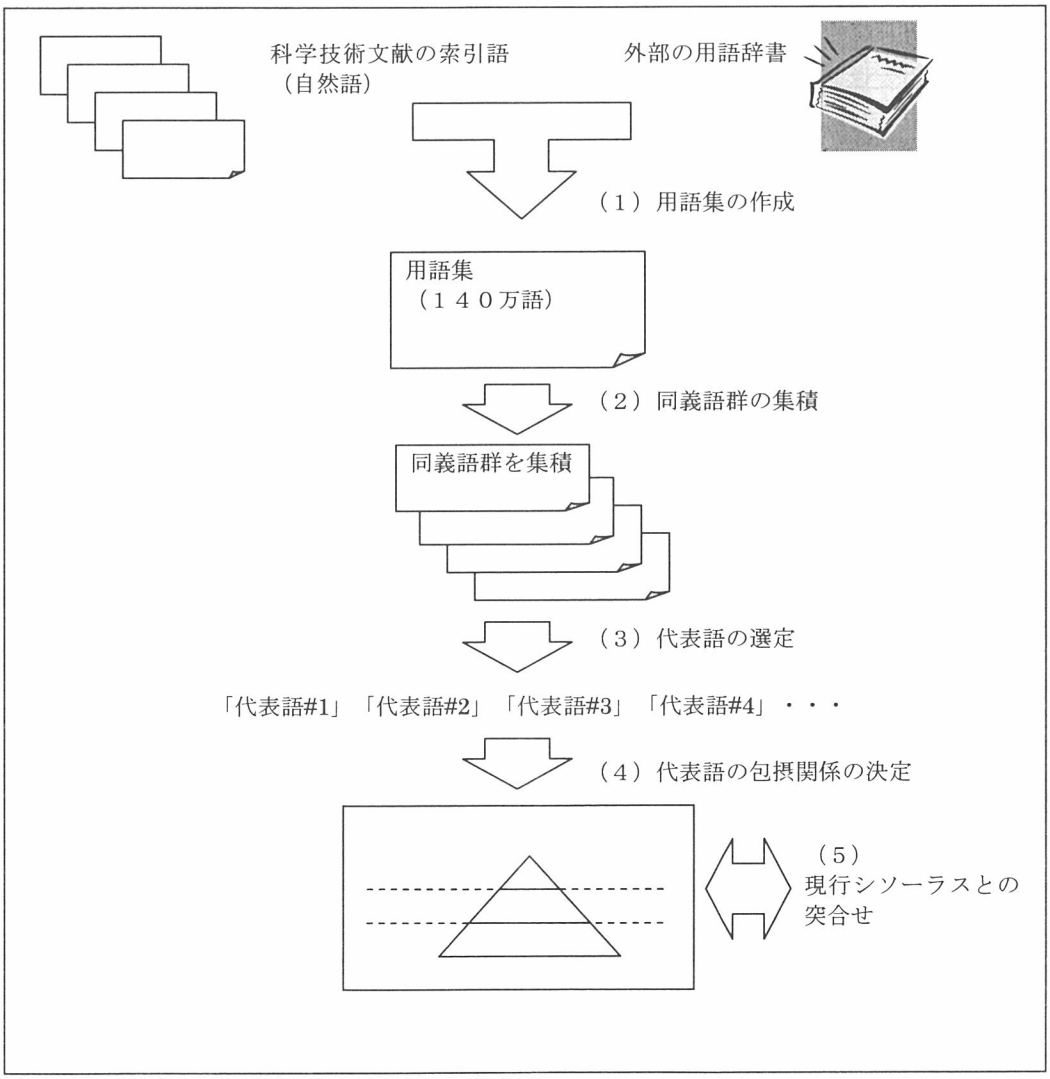
（１）科学技術文献にはシソーラス用語以外に自然語で索引付けを行っている．これに外部の辞書の用語を付け加え，まず，約 140 万語の用語集を作成した．

（２）分野別の用語辞書や和英辞書を用いて同義関係を判断し，同義語群の集積を行った．また，「概念」ではない用語を削除し，約 42 万語まで圧縮した．約 42 万語に対してシソーラス用語との包摂関係，同義語関係を付与したものが現在の大規模用語辞書である．

（３）シソーラスの改訂にあたっては，現在付与されているシソーラス用語から一旦離れ，大規模用語辞書の同義語群ごとに索引頻度等を参考にしつつ新たに代表語を選定した．

（４）現在，代表語間の包摂関係を整理し、階層情報つきの代表語を作成している．

（５）今後，階層情報つきの代表語を元に，現行のシソーラス用語の追加，削除を実施する予定である．



（図９）大規模辞書の構築とシソーラスの改訂手順

シソーラスの改訂作業において、最も重要なステップは「（３）新たな代表語の選定」と「（４）その包摂関係の決定」である．

JST では，図１０のようなシートを作成し，集積した同義語群毎に，代表語を選定している．代表語の選定を効率的に実施できるよう，シートには索引頻度や用語カテゴリ（JST 分類コード）も記載している．

また，階層構造の付与については，JST では KWIC（KeyWord In Context）を主に使用している．分野によっては，カテゴリ分割のタグ（例えば UDC）を各用語に付与することにより，階層関係を構築していくことも可能と思われる。

索引頻度合計	同義語ID	準ディス	出典	索引頻度	第１主題	第２主題	第３主題	関係子1	対応ディスク	関係子2	対応ディスク	関係子3
13921	0019582	看護師		8575	LS52			SY	看護婦			
13921	0019582	看護士		5346	LS52			SY	看護婦			
7567	0115303	ＥＢＭ		6695	LS52			BT	治療計画	RT	治療法	
7567	0115303	エビデンス		813	LS52			BT	治療計画	RT	治療法	
7567	0115303	根拠に基づく医療		59				BT	治療計画	RT	治療法	
7376	0016207	細胞移植		7098	LS52			BT	移植【動物】			
7376	0016207	細胞療法		162	LS52			BT	移植【動物】			
7376	0016207	細胞治療		116	LS52			BT	移植【動物】			
6887	0020183	クリニカルパス		5402	LS52			BT	治療計画			
6887	0020183	クリティカルパス		1387	LS52			BT	治療計画			
6887	0020183	クリティカル・パス		52	LS52			BT	治療計画			
6887	0020183	クリニカル・パス		46	LS52			BT	治療計画			
5992	0017043	病院管理		5825	LS53			BT	管理			
5992	0017043	病院経営		167	LS53			BT	管理			
5393	6	リンパ節転移		5393								
4751	0015044	大腸内視鏡検査		4751	LS52			SY	結腸鏡検査			
4628	0018247	遺伝子多型		4497	LS18			BT	変異	RT	遺伝子型	
4628	0018247	遺伝的多型		74	LS18			BT	変異	RT	遺伝子型	
4628	0018247	遺伝子多形		57	LS18			BT	変異	RT	遺伝子型	
4549	0102829	生活習慣病		4549	LS51			BT	成人病			
4365	0006845	慢性疾患		4365	LS51			BT	病気			
4145	0016809	統合失調症		4145	LS51			SY	精神分裂病			
4089	0006384	皮疹		4089	LS51			SY	発疹			
4043	0014718	胸腔鏡		4043	LS54			BT	内視鏡			
3967	0095023	母親		3967	LS14			RT	母性			
3680	0014499	ＲＣＴ		3680	LS41			SY	無作為化比較試験			
3411	0017103	再生医療		2120	LS52			RT	再生	BT	医学	
3411	0017103	組織工学		543	LS52			RT	再生	BT	医学	
3411	0017103	再生医学		536	LS52			RT	再生	BT	医学	
3411	0017103	組織再生		97	LS52			RT	再生	BT	医学	
3411	0017103	ティッシュエンジニアリング		77	LS52			RT	再生	BT	医学	
3411	0017103	生体組織工学		38	LS52			RT	再生	BT	医学	
3271	0001420	心筋細胞		3271	LS17			BT	細胞			
3235	0090156	ＣＢＤＣＡ		2079	LS44			BT	抗腫よう薬	BT	白金錯体	
3235	0090156	カルボプラチン		1156	LS44			BT	抗腫よう薬	BT	白金錯体	
3164	0006843	急性疾患		3164	LS51			BT	病気			
3086	0010910	マトリックスメタロプロテイナーゼ		3034	LS38			BT	メタロプロテイナーゼ			
3086	0010910	ＭＭＰ		52	LS38			BT	メタロプロテイナーゼ			
3050	0115477	電子カルテ		3050	LS52			RT	病歴	BT	記録	
2973	0007759	腸性リンパ腫		2973	LS51			SY	リンパ腫			
2934	0017336	保健教育		2536	LS55			RT	健康管理			
2934	0017336	健康教育		227	LS55			RT	健康管理			
2934	0017336	保健指導		171	LS55			RT	健康管理			
2901	0093063	穿刺吸引細胞診		2901	LS52			BT	細胞診			
2760	0010484	ＢＮＰ【ペプチド】		2270	LS33	LS37		BT	生理活性ペプチド			
2760	0010484	脳性ナトリウム利尿ペプチド		336	LS33	LS37		BT	生理活性ペプチド			
2760	0010484	脳ナトリウム利尿ペプチド		154	LS33	LS37		BT	生理活性ペプチド			

（図１０）集積された同義語群と代表語決定用シート

５．あとがき

JST では，ユーザの要望を受け，各種の機能改善を実施している．一方，JST 自ら広い視点で検索システムのあるべき姿を研究し，より良いサービスを提供していきたいと考えている．

科学技術文献のデータベースに比べ，人文学のデータベースはデータの形態は多様であり，また概念同士の関連も複雑と思われる．科学技術分野と人文学分野のデータベース作成，提供の経験が交流することで互いの発展につながることを期待したい．

参考文献

[1]新井兼,甲田彰:新文献検索システム「JDream Ⅱ」の開発,情報メディア学会研究大会発表資料,Vol.5, Page.29-32,2006