

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「JDream II」の提供を開始した。JDream IIには様々な検索支援機能を搭載しており、そのうちのひと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、JDream II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「JDream II」の提供を開始した。JDream IIには様々な検索支援機能を搭載しており、そのうちのひと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、JDream II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. JDream IIの概要

2. 1. JDream II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JDream II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

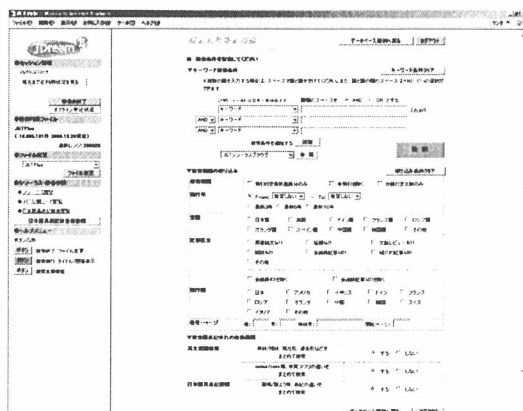
(表 2) 他機関のデータ

2. 2. JDream II の概要

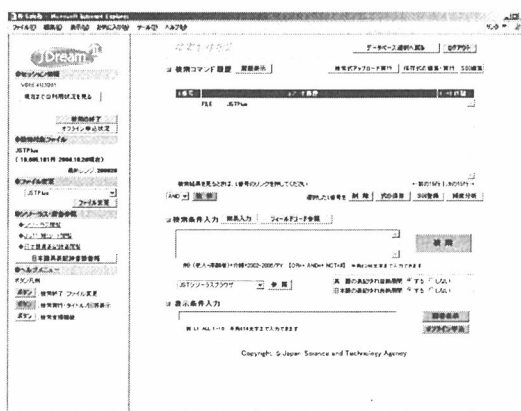
JDream II は上記の文献データベースを検索するための WEB ベースのシステムであり, 我が国の科学技術文献検索サービスで最も利用されているシステムである.

JDream II 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.

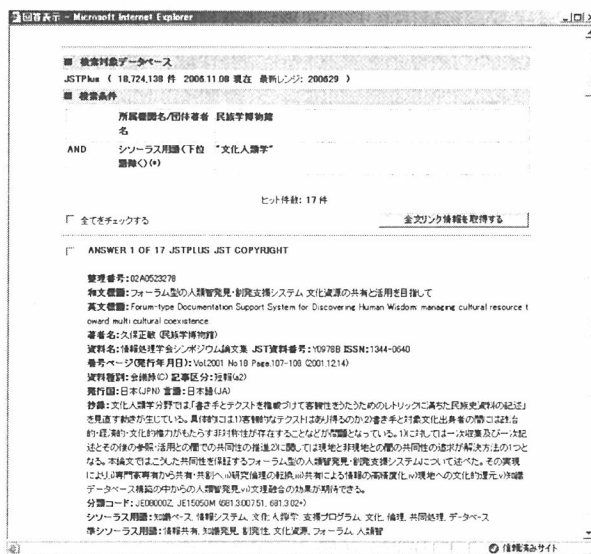
シンプルモードは検索の結果に対して検索条件の追加, 頻度分析結果による絞り込み検索が可能で, 初心者でも迷わず使用できる「直線的」なインターフェースである (図 1). 一方, コマンドモードは, 検索結果の集合間の論理演算や検索式の保存・再利用が可能な「試行錯誤型」のインターフェースである (図 2). いずれのモードでも科学技術文献情報の表示 (図 3) は共通で, ユーザの習熟度や利用状況に応じた使用が可能である.



(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. JDream II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。

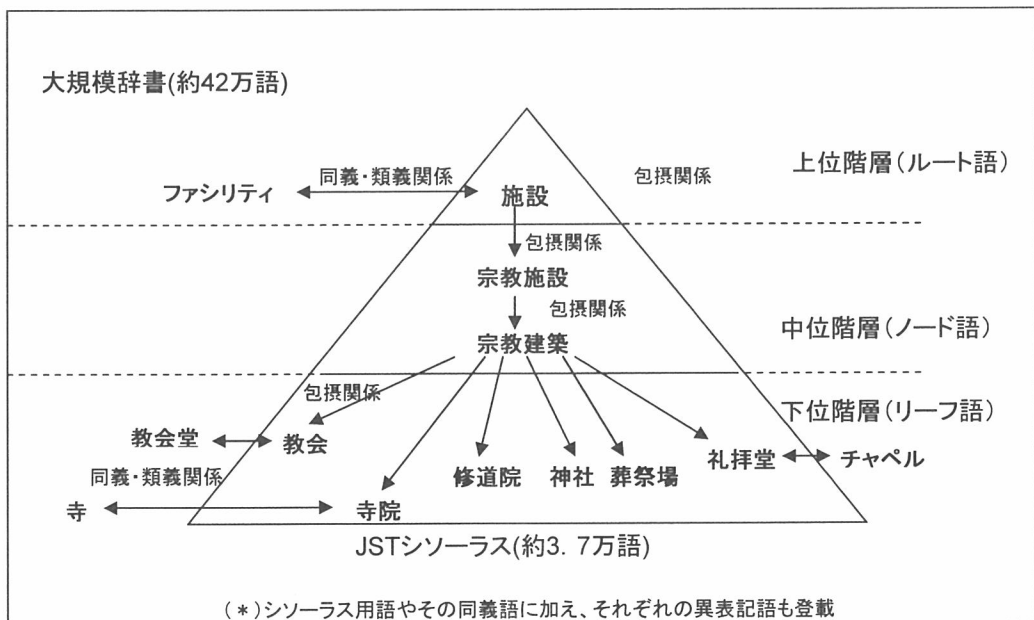
これに対して、JDream II では、ユーザがより簡便に索引語を見つけだすことができるよう、科学技術関連用語の同義語や異表記語を収録した大規模用語辞書と、それを JDream II のユーザ

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、JDream II 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, 禮拜堂, …

(表 2) 人文学系の JST シソーラス用語 (抜粋)

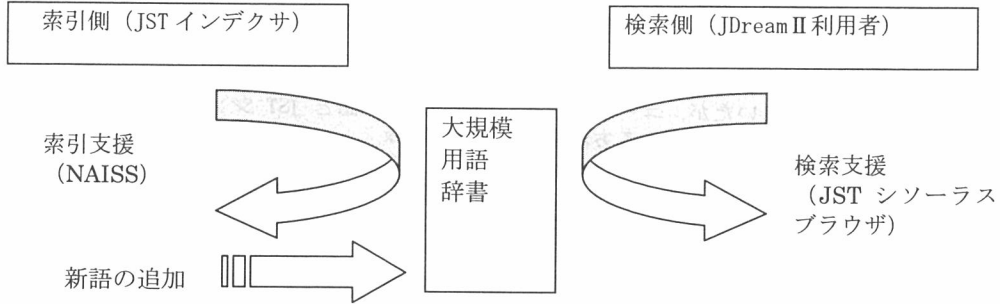
3. 3. 大規模用語辞書の特徴

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 の検索で使うことが可能である[1]。



(図5)索引側と検索側からみた大規模用語辞書

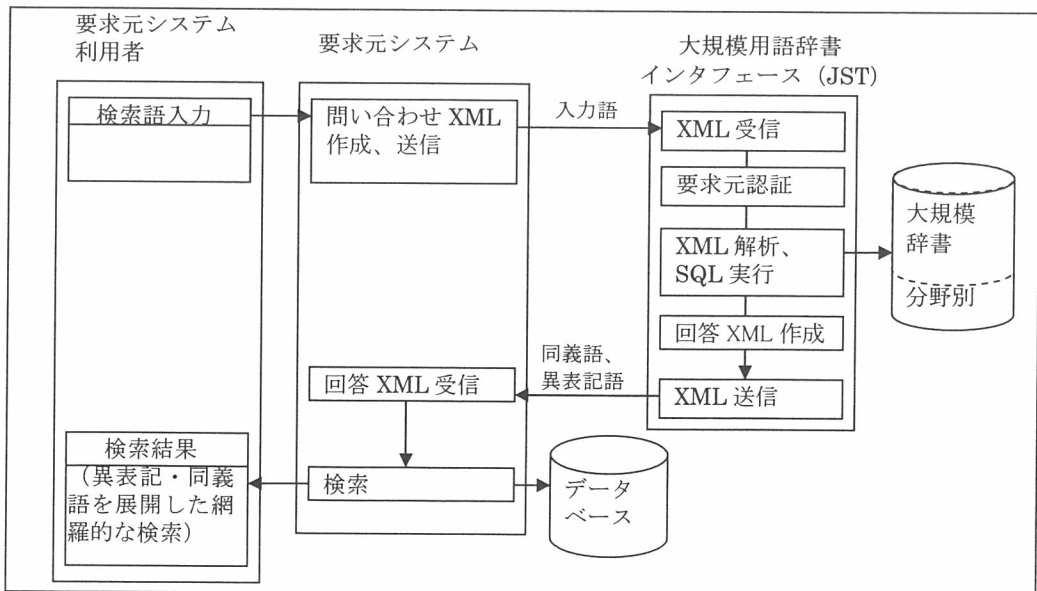
4. 人文学分野への大規模用語辞書の応用

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、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。

The screenshot displays the NAISS interface with the following elements:

- Document Header:** Includes fields for '抄録・索引' (Abstract/Index), '受入番号' (20050781682), '記事番号' (0001), and '入力途中' (Input in progress).
- Main Text Area:** Contains document content with a highlighted sentence: '用語を反転表示させクリックすると、その用語が「索引候補語」欄に自動的に入力' (Clicking the term to be displayed in reverse will automatically input the term into the 'Index Candidate Words' field).
- Search Results Panel:** Shows a list of search results, including 'プレッシャー' (Pressure) and 'プレッシャーイング吸' (Pressure-ing suction).
- Annotations:**
 - A callout box points to the search results, stating: '登録語リストから用語を選択すると、索引に適したソーラス用語を表示' (Selecting a term from the registered word list displays terminology suitable for indexing).
 - Another callout box points to the search results, stating: '自然語をもとに、大規模辞書登録語を検索し、結果を表示' (Search for registered words in the large dictionary based on natural language and display the results).
 - A large cylinder icon labeled '大規模用語辞書' (Large Terminology Dictionary) is connected to the search results area by a dashed line.
- Footer:** Includes navigation buttons like '検索' (Search), '表示' (Display), and '物理支援' (Physical Support).

(図 7)索引支援システム (NAISS) のインタフェース例

JDream II シソーラス辞書参照機能 - 自然語検索画面 - Microsoft Internet Explorer

JDream II 閉じる

■ 自然語から索引語を見つける

JSTでは独自に作成しているシソーラスの用語を用いて各文庫に索引しています。
索引語の種別としてはシソーラス用語、準シソーラス用語、化学物質名があります。
JSTシソーラスブラウザでは辞書から索引語とその同義語、具表記語を検索することができます。

ドラマ で始まる 語を辞書から検索

■ 候補語一覧

「ドラマ」で始まる 自然語をもとに、大規模辞書登録語を検索し、結果を表示

#	ヒットした語	索引語、bdf	種別	詳細を
1	ドラマ	[演劇 の同義語]	シソーラス	表示
2	ドラマミン	[J252.867C]	化学物質	表示
3	ドラマリン	[J252.867C]	化学物質	表示
4	ドラマンド戯候	[Drummond戯候 の具表記語]	準シソーラス	表示

ページが表示されました 情報読み込み

JDream II シソーラス辞書参照機能 - 索引語詳細画面 - Microsoft Internet Explorer

JDream II 閉じる

[候補一覧に戻る](#)

索引語	演劇
英語表記	drama
種別	シソーラス用語

[シソーラス階層を表示](#)

検索語入力エリアに反映

同義語で検索範囲を広げる サブヘディングで検索範囲を絞り込む

下記の同義語を検索に加える

以下をすべて選択

ドラマ

プレイ

劇

芝居

検索語入力エリアに反映

「演劇」に關係しているシソーラス用語
クリックするとその語の索引語画面に切り替わります

直近上位語群	芸術
直近下位語群	
関連語群	劇映画 娯楽

ページが表示されました 情報読み込み

大規模用語辞書

(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。

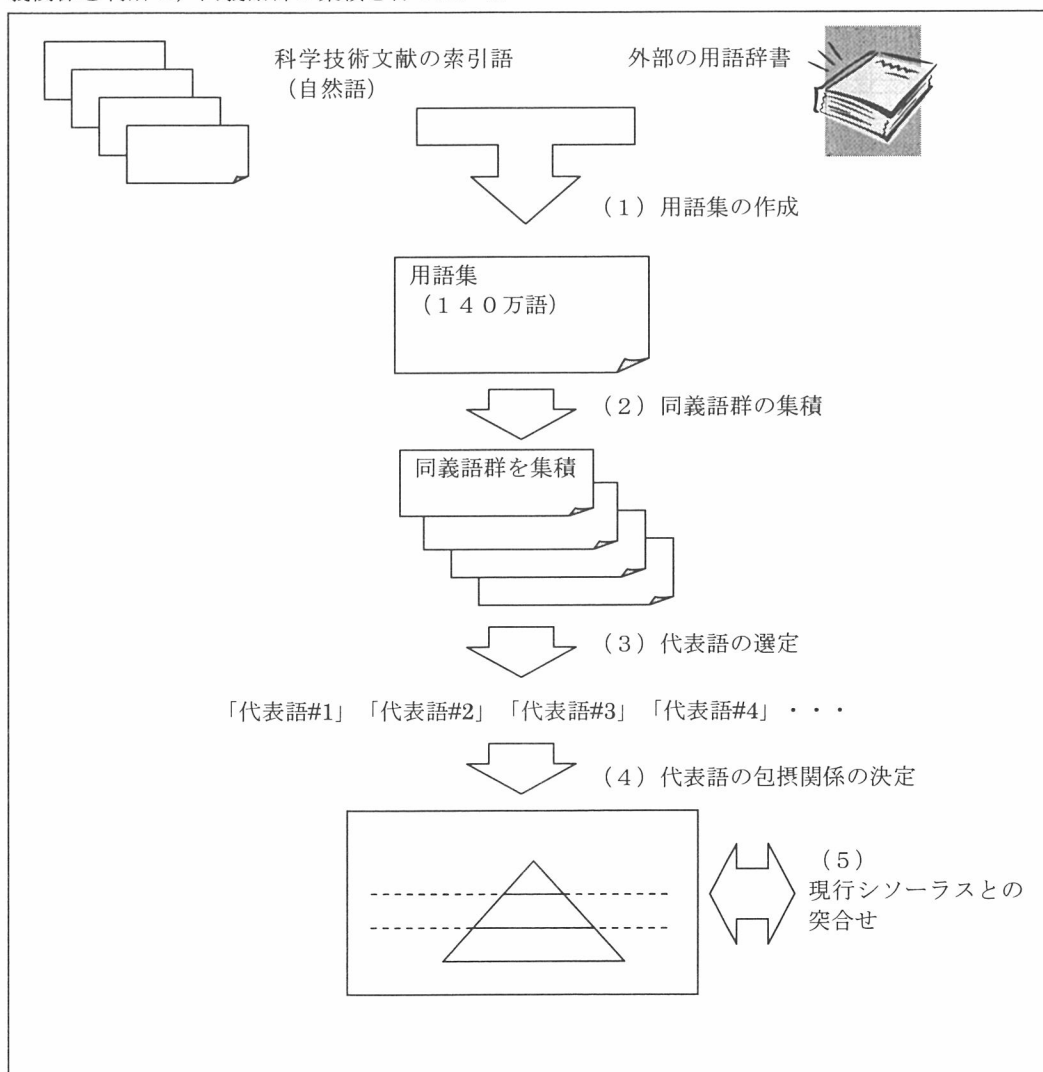
(2) 分野別の用語辞書や和英辞書を用いて同義関係を判断し、同義語群の集積を行った。ま

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	同義語ID	準ディク	出典	索引頻度	第1主題	第2主題	第3主題	関係子1	対応ディク	関係子2	対応ディク	関係子3
13921	0019582	看護師		8676	LS52			SY	看護職			
13921	0019582	看護士		5346	LS52			SY	看護職			
7567	0115303	E B M		6695	LS52			BT	治療計画	RT	治療法	
7567	0115303	エビデンス		913	LS52			BT	治療計画	RT	治療法	
7567	0115303	根拠に基づいた医療		59				BT	治療計画	RT	治療法	
7376	0018207	細胞移植		7098	LS52			BT	移植【動物】			
7376	0018207	細胞療法		162	LS52			BT	移植【動物】			
7376	0018207	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			
5932	0017043	病棟管理		5825	LS53			BT	管理			
5932	0017043	病棟運営		167	LS53			BT	管理			
5939	0	リンパ節腫瘍		5333								
4751	0015044	大腸内視鏡検査		4751	LS52			SY	結腸直腸検査			
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	遺伝子多型		57	LS18			BT	変異	RT	遺伝子型	
4548	0103299	身障者酒		4548	LS51			BT	成人酒			
4365	0006845	慢性疾患		4365	LS51			BT	成人酒			
4145	0018809	統合失調症		4145	LS51			SY	精神分裂病			
4089	0006384	皮膚		4089	LS51			SY	皮膚			
4043	0014718	胸腔鏡		4043	LS54			BT	内視鏡			
3887	0095029	母乳		3887	LS14			RT	母乳			
3680	0014499	R C T		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		37	LS52			RT	再生	BT	医学	
3411	0017103	メ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	C B D C A		2079	LS44			BT	抗腫瘍薬	BT	白金剤	
3235	0090156	カルボプラチン		1156	LS44			BT	抗腫瘍薬	BT	白金剤	
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ナーゼ		3034	LS38			BT	メタロプロテナーゼ			
3086	0010910	MMP		52	LS38			BT	メタロプロテナーゼ			
3050	0015472	電子カルテ		3050	LS52			RT	病歴	BT	記録	
2973	0007759	属性リンパ球		2973	LS51			SY	リンパ球			
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	疫学疫学		2901	LS52			BT	細胞診			
2760	0010484	B N P 【ペプチド】		2270	LS39		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	LS39		LS37	BT	生理活性ペプチド			
2760	0010484	脳ナトリウム利尿ペプチド		164	LS39		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