WoSCC-NISTEP 大学・公的機関名辞書対応テーブル 説明書

2018 年 10 月

文部科学省科学技術・学術政策研究所

1. はじめに

研究論文等のデータベースの利用に際して、機関名で検索したり、機関別の集計や分析を行ったりすることがよくあります。そのときの厄介な問題の一つは、機関名の表記が統一されておらず、いろいろな「表記のゆれ」が見られることです。英語のデータベースで、たとえば東京農工大学の正式英語名は Tokyo University of Agriculture and Technology ですが、これが Tokyo Noko University、Tokyo Agriculture and Technology University などと表記されたり、”University”が”Univ”、” Agriculture and Technology”が”A&T”などと略記されたりします。

この問題は、データベースに含まれる機関名データがどの機関を表しているかを正しく同定できれば解決されます。科学技術・学術政策研究所(NISTEP)では、世界最大級の書誌・引用データベースである Web of Science Core Collection（以下 WoSCC と略）に含まれる機関名データから、国内の機関（どのような機関を含むかについては 2. (2)をご覧下さい）の機関同定を行っています。この結果に基づき、WoSCC の機関名データを、NISTEP 大学・公的機関名辞書（以下、「機関名辞書」）の収録機関に対応させる「WoSCC-NISTEP 大学・公的機関名辞書対応テーブル」（以下、「このテーブル」と呼ぶ）を作成しました。WoSCC データベースの利用や、国内機関の論文生産に関する調査分析に役立てていただくことを念頭に、WoSCC の提供元である Clarivate Analytics(旧トムソン・ロイター IP&Science)の了解を得て、このテーブルを公開いたします。

なお、このサイトから既に公開している以下のデータも併せてご利用下さい。

- ○ NISTEP 大学・公的機関名辞書データ：約 19,000 の国内機関の和英の名称、属するセクター、変遷情報（統廃合、改称等）等を収録した辞書データです。大学、公的機関が中心ですが、研究活動を行っているそれ以外の機関もできるだけ収録しています。このテーブルにおける機関同定は、この辞書に基づいています。（説明資料：NISTEP 大学・公的機関名辞書利用マニュアル）
- ○ 大学・公的機関名英語表記ゆれテーブル：機関名辞書に含まれている機関の英語表記（正式名の他、通称、略称等の別名を含む）と、WoSCC データベースに現れる主な機関名英語表記のデータを統合ししたデータです。（説明資料：大学・公的機関名英語表記ゆれテーブル利用の手引き）

※ このテーブルの改訂について

「WoSCC-NISTEP 大学・公的機関名辞書対応テーブル(ver.2017.1)」（2017 年 4 月公開）は、2015 年末時点における WoSCC カスタムデータから抽出した機関名データを機関名辞書 Ver.2016.1（2016 年 10 月公開）を用いて同定したものです。その後の機関名辞書の更新のため、同定された機関の識別番号(ID)や名称が、現在の機関名辞書に存在しないものが若干生

じました(全レコードの 0.2%程度)。その主な理由は、日本語名称を正確な名称に修正したため、もとの機関エントリーを削除して正しい名称で再登録したためです[1]。

今回(2018 年 9 月)、公開している最新の機関名辞書(ver.2018.2)に合わせて、それらの同定先を更新し、「WoSCC-NISTEP 大学・公的機関名辞書対応テーブル(ver.2017.1.1)」としました。WoSCC のデータには変更はありません。

※ このテーブルのデータ分析への利用について

このテーブルは、Clarivate Analytics との WoSCC の利用ライセンス契約により、NISTEP が WoSCC の二次的著作物として作成したものです。従って、NISTEP と Clarivate Analytics の両者が著作者です。テーブル中の WoSCC 記事番号(WoS_ut)、論文出版年(bib_date_year)、著者所属機関番号(rs_address_seq)は WoSCC から抽出したデータ、他は NISTEP が作成したデータです(詳しくは 3.をお読みください)。

このテーブルを用いたデータの分析、及び分析結果の公表は、下記によるものとします。

(1) データ分析への使用は自由です。分析に必要なデータの複製も、外部に公表されない限り自由です。

(2) 但し、このテーブル(データ)を WoSCC と組み合わせて利用される際には、Clarivate Analytics との契約に従ってください。

(3) 分析の結果を本テーブルの二次的著作物として公表される場合、次のように原著作者のクレジットを表示してください。

> 原著作者名: 文部科学省科学技術・学術政策研究所(NISTEP)
> Copyright 2018 - Clarivate Analytics. All rights reserved.
> 作品タイトル: WoSCC-NISTEP 大学・公的機関名辞書対応テーブル
> URL:http://www.nistep.go.jp/research/scisip/data-and-information-infrastructure

(4) 以上の利用には、営利目的の利用を含むものとします。

(5) データ分析以外の本テーブルの利用(複製、公衆送信等)については、著作者とご相談ください。

2. 同定の対象と方法

(1) 同定対象のデータ

今回同定を行ったデータは、WoSCC データベースに採録された論文の著者所属機関データのうち、下記の条件に当てはまるものです。該当の論文は約 134 万件、その中の日本機関のデータは延べ 273 万件です。

(a) Science Citation Index Expanded に収録された論文

[1] 機関名辞書では、もとの機関自身の名称が変更された場合には、旧機関を削除することはせず、新機関との間に関連づけをしますが、機関の日本語正式名を訂正した場合(半角文字を全角文字に変更した場合なども含む)は、訂正前の機関エントリーを削除して新しいエントリーに入れ替えます。

(b) 論文出版年が 1998～2015 年

(c) ドキュメントタイプが"article"または"review"

(d) 日本の機関と判別された著者所属機関データ(著者所属機関所属国が"Japan")

(2) 同定の方法

日本の機関と判別された著者所属機関データを、個々に機関名辞書に収録されている英語名称(正式名の他、通称、略称等の別名を含む)と照合することにより、同定を行います。

機関名辞書には、独立した機関(これを代表機関と呼んでいます)の他、代表機関に属する主要な下部組織も収録しています(約 19,300 機関中 3,700 機関が下部組織です)。特に、論文数の多い 32 大学については重点的に下部組織を収録しています。代表機関とその下部組織がともに同定された場合は、下部組織が優先されます。なお、機関名辞書における代表機関の考え方については、「NISTEP 大学・公的機関名辞書利用マニュアル」をご覧ください。

また、機関名辞書では、機関を 16 のセクターに分類しています(3.(f)を参照)。これらのセクターには、大学や公的機関の他、地方公共団体の機関、会社、非営利団体等も含まれていますので、それらに属する機関も同定の対象になります。

(3) 同定フラグ

同定のレベルを 5 段階で区分します。WoSCC の各機関名データに対し、次の順序でマッチングを行い、同定します。同定フラグが S, H, N のデータは、機関同定ができなかったものです。

同定フラグ	説明
L	WoSCC 機関表記に最長マッチした機関名辞書の機関に同定。
M	曖昧マッチング(N-gram とレーベンシュタイン距離を使用したマッチング)と郵便番号マッチングの結果が一致した場合、その機関に同定。
S	機関同定ができなかったがセクターが同定できたデータ。
H	機関もセクターも同定できなかった病院であることが同定できたデータ。
N	国内機関であることのみ同定できたデータ。

3. テーブルの構成

このテーブルは、論文の発表年(4. (b)の"bib_date_year")により、以下の 3 つの tsv ファイルに分離されています。

WoS_NID_corres_1998_2004_ver.2017.1.1.tsv: 論文発表年が 1998～2004 年のデータ

WoS_NID_corres_2005_2010_ver.2017.1.1.tsv: 論文発表年が 2005～2010 年のデータ

WoS_NID_corres_2011_2015_ver.2017.1.1.tsv: 論文発表年が 2011～2015 年のデータ

各ファイルのデータ形式は全く同じです。

4. テーブルの各項目

テーブルの各項目について説明します。

(a) WoSCC 記事番号(WoS_ut):当該機関を著者所属機関に含む WoSCC の記事番号です。

(b) 論文出版年(bib_date_year):論文が発表された年です。

(c) WoSCC 記事内の著者所属機関番号(rs_address_seq):1 つの WoSCC 記事(WoS_ut が同一)の中に存在する著者所属機関レコードの中での当該レコードの順番です。最初のレコード番号が 1、以下 2,3,…となります。日本以外の所属機関のレコードはこのテーブルに含まれていませんので、同じ WoS_ut の中で番号が飛んでいることがあります。

(d) 同定フラグ:2.(3)で述べた L, M, S, H, N のいずれかです。同定フラグが S のレコードでは以下の(e), (g), (h)が、H または N のレコードでは以下の(e), (f), (g), (h)が空白です。

(e) 機関 ID:同定された機関に機関名辞書で与えられている識別番号です。この番号を用いて、機関名辞書により機関の英語名称、上位機関、変遷等の情報を得ることができます。詳しくは「NISTEP 大学・公的機関名辞書利用マニュアル」をご覧ください。

(f) セクター番号とセクター分類:同定された機関が属するセクターです。機関名辞書では、次の表に示すように、機関を 16 のセクターに分類しています[2]。

	セクター番号	セクター分類
大学等	1	国立大学
	2	国立短期大学
	3	国立高等専門学校
	4	公立大学
	5	公立短期大学
	6	公立高等専門学校
	7	大学共同利用機関
	12	私立大学
	13	私立短期大学
	14	私立高等専門学校
公的機関	8	国の機関
	9	国立研究開発法人等*1
その他の機関	10	地方自治体の機関*2
	15	会社
	16	非営利団体
	17	その他の機関

*1 独立行政法人、特殊法人、認可法人を含む。

*2 地方独立行政法人を含む。

[2] この他に学校法人(セクター番号 11)がありますが、機関同定には使用していません。

(g) 機関正式名:同定された機関の日本語正式名です。

(h) 代表機関名:同定された機関が属する最上位の機関(2.(2)で述べた代表機関)です。同定された機関が下部組織の場合はその代表機関名を、代表機関の場合は代表機関名自体を記載しています。代表機関の場合は空欄としてもよいのですが、配列や集計に便利なように、このような記載としました。

(i) 同定番号:一つのWoSCC機関データが複数の機関に同定されることがあります。たとえば、"National Institute of Genetics, The Graduate University for Advanced Studies (SOKENDAI)"という例では、国立遺伝学研究所と総合研究大学院大学という 2 つの異なる機関が 1 つの機関名レコードに記載されています(このような例は、主に一人の著者が異なる機関に属する場合に見られます)。このような場合、このテーブルでは複数の同定機関を別々のレコードに分割し、それらの同定番号をそれぞれ 1, 2 として区別します。WoS_ut と rs_address_seq は同じになります。

(j) 同定数:上記の同定番号の繰り返し数です。このテーブルでは、WoSCC の全所属機関データ中、同定数 1 が 99.4%で、残りが同定数 2～4 です。

5. 同定結果の概要

セクターごとの同定フラグの分布は次の通りです。同定数が 2 以上の場合、それぞれを独立してカウントしています。このため、合計数は2(1)で述べたもとのWoSデータ数よりやや多くなっています。機関同定されたデータ(同定フラグがLまたはM)は、全体の94.6%です。また、機関同定されたうちでは、大学等 74.6%、公的機関 13.6%、その他の機関 11.8%となります。

セクター	同定フラグ					
	L	M	S	H	N	計
国立大学	1,319,347	159	0	-	-	1,319,506
国立短期大学	181	0	0	-	-	181
国立高等専門学校	8,907	52	0	-	-	8,959
公立大学	124,667	62	0	-	-	124,729
公立短期大学	700	0	0	-	-	700
公立高等専門学校	796	1	0	-	-	797
大学共同利用機関	38,532	0	0	-	-	38,532
私立大学	445,856	464	0	-	-	446,320
私立短期大学	2,065	0	0	-	-	2,065
私立高等専門学校	55	1	0	-	-	56
国の機関	57,212	31	122	-	-	57,365
国立研究開発法人等	296,125	416	0	-	-	296,541
地方自治体の機関	67,410	85	9,915	-	-	77,410
会社	160,686	0	37,541	-	-	198,227
非営利団体	77,891	175	0	-	-	78,066

その他の機関	593	0	0	-	-	593
不明	-	-	-	52,781	48,731	101,512
計	2,601,023	1,446	47,578	52,781	48,731	2,751,559

6. このテーブルの利用法

このテーブルは、主に次の 2 つの利用法が考えられます。

(1) WoSCC での著者所属機関検索・分析の補助ツールとして

これには次の二通りの利用が考えられます。なお、1.で述べたように、WoSCC を利用するには、Clarivate Analytics との契約が必要です。

第一は、WoSCC で検索した論文データ集合における所属機関(大学または公的機関)の同定(名寄せ)です。WoSCC のカスタムデータを用いる場合は、このデータ中の ut と rs_address_seq の項目を、このテーブルの WoS_ut 及び rs_address_seq と接合することで、機関名の名寄せが可能となります。WoSCC のオンラインデータを用いる場合は、検索結果をダウンロードしたファイルを用います。ダウンロードデータでは、WoS_ut は UT の項目にあります。rs_address_seq に相当する項目はありませんが、著者所属機関を示す C1 項目中に配列されている順番がその番号に相当します。

第二の利用方法は、ある機関の論文データの一括検索です。まず、検索したい機関の機関 ID を機関名辞書で調べます。次に、このテーブルを用いてその機関 ID を持つ論文データに対する WoS_ut の集合を作り、WoSCC データベースからそれらに一致するレコードを抽出します。これにより、WoSCC 中の機関名表記のゆれに関わりなく、漏れのない機関検索が行えます。WoSCC のカスタムデータには、この方法を直接適用できます。オンラインデータを用いる場合は、(1)と同様に検索結果をダウンロードします。ダウンロードしたファイルの各レコードに WoS_ut が付けられていますので、この方法が適用できます。別の方法として、このサイトで公開している「大学・公的機関名英語表記ゆれテーブル」によって検索したい機関の表記バリエーションを取得し、それらを用いて機関名の OR 検索を行うこともできます。

(2) 国内機関の論文生産統計の基礎データとして

このテーブルと機関名辞書を用いて、1998-2015 年の期間における機関の論文生産統計をとることができます。代表機関別の統計、セクター別の統計も得ることができます。

但し、レコードを単純に集計した結果は、機関またはセクターの合計論文数ではなく、WoSCC データベースに出現した著者所属機関レコードの合計数であることにご注意下さい。一つの論文に同じ機関の異なる部局の著者が含まれている場合、この機関のレコードが複数存在する(それぞれ部局が異なる)ことがあります。論文数の統計をとる場合には、同じ WoS_ut の中の同じ機関(機関 ID が同じ)のレコードの重複を削除する必要があります。

WoS_ut を用いると、異なる機関あるいは異なるセクターの間でどれくらい共著論文があるか(共同研究が行われているか)を調べることもできます。

なお、このテーブルで可能なのは、1998-2015 の期間にわたる統計だけです。期間、分野、ドキュメントの種類を区切った統計を得るには、WoSCC データベースと情報を組み合わせる必要があります。

7. 注記

(1) WoSCC-NISTEP 大学・公的機関名辞書対応テーブルの精度

このテーブルの作成には十分な注意を払っておりますが、すべての同定結果を人手でチェックはしていませんので、少数の同定エラーがあります。サンプルデータのチェックの結果では、機関同定できたデータ(同定フラグが L または M)のうちエラー率は 0.3%未満です。その内訳は、代表機関が間違っていたもの 0.06%、代表機関は正しく同定されたが下部組織が間違っていたもの 0.03%、複数同定(3.(k)に述べた同定数が 2 以上)の一方のみが正しかったもの(単一同定とすべきもの)が 0.20%です。

下部組織の同定結果については、組織名の表記ゆれや NISTEP 大学・公的機関名辞書に収録されていない組織などの影響で、同定率や精度が代表機関と比べて低くなっています。このテーブルの活用の用途に応じて、目視確認等をお願いします。

同定アルゴリズムの精密化、機関名辞書のデータ充実等により更に改善を行っていく予定ですが、ご使用に当たって注意下さるとともに、お気づきの点をお知らせ下さい。

(2) このテーブルのカバー率

このテーブルのデータは、2015 年末時点における出版年 1998 年以降の WoSCC カスタムデータから抽出しました。しかし、WoSCC では、適時データの追加、修正が行われていることから、1998～2015 年についても、カバー率は 100%とはなっていません。2019 年 9 月 3 日時点の WoSCC に含まれている日本論文(2(1)の条件(a)～(c)に当てはまり、条件(d)に当てはまるデータを少なくとも 1 件含む論文)の数と、このテーブルがカバーする論文数の比較を下表に示します。2015 年のカバー率がやや低いのは、2015 年末時点にデータを抽出したためです。

出版年	WoSCC 日本論文数(2018.9.3 時点)	このテーブルに含まれる論文数	カバー率
1998	69,313	68,753	99.2%
1999	71,352	70,856	99.3%
2000	73,544	73,001	99.3%
2001	73,057	72,563	99.3%
2002	74,435	73,960	99.4%
2003	76,656	76,096	99.3%
2004	77,081	76,708	99.5%
2005	76,809	76,458	99.5%
2006	77,228	76,817	99.5%
2007	75,832	75,443	99.5%

2008	76,205	75,907	99.6%
2009	75,560	75,163	99.5%
2010	74,435	74,123	99.6%
2011	76,421	76,045	99.5%
2012	77,113	76,787	99.6%
2013	78,615	78,238	99.5%
2014	77,219	76,730	99.4%
2015	76,624	69,323	90.5%
計	1,357,499	1,342,971	98.9%