WoSCC-NISTEP 大学・公的機関名辞書対応テーブル 説明書

2021 年 1 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 の提供元であるクラリベイト社の了解を得て、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 20,000 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 及び Scopus 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

これまで公開していた「WoSCC-NISTEP 大学・公的機関名辞書対応テーブル(ver.2017.1.2)」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.2018.2（2018 年 8 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によっ

て、若干の機関データ(機関 ID と日本語名称)の入れ替えを行っています)[1]。

今回、2019 年 12 月末時点における WoSCC カスタムデータから抽出した機関名データを、最新版の機関名辞書 Ver.2020.1(2020 年 6 月公開)を用いて新たに同定を行い、「WoSCC-NISTEP 大学・公的機関名辞書対応テーブル(ver.2020.1)」として公開しました。

※ 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、クラリベイト社との 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 とクラリベイト社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)、記事内アドレス番号(rs_address_seq)、記事内アドレス数(rs_address_count)、機関・アドレス情報(rs_organization、rs_suborganization、rs_address)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 4.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、クラリベイト社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 2020 - 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://doi.org/10.15108/data_rsorg003_2020_1

(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 168 万件、その中の日本機関のデータは延べ約 358 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 1998～2019 年(同定に用いた WoSCC には 2020 年出版の論文も多少含まれていますが、これは含めていません。)

(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")

これまでの版(Ver.2017.1.2)では論文数約 134 万件、日本機関データ延べ約 237 万件でした。データの増は対象の論文出版年を 2015 年までから 2019 年までに延ばしたことによります。

(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 20,000 機関中 4,000 機関が下部組織です)。特に、論文数の多い 32 大学及び下部組織収録の協力をいただいた 1 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(4.(e)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。
R	たとえば、農業・食品産業技術総合研究機構花き研究所、自然科学研究機構核融合科学研究所、鹿屋体育大学はいずれも NIFS という英語名(略名)を持つ。このような場合、所在地、郵便番号等の情報から適切と考えられる機関に同定。
V	L, M, R のいずれによっても同定機関が決まらない場合、機関名データと辞書の名称データのベクトル類似度がある閾値以上の最高値を示す機関に同定。
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、論文の出版年（4. (b)の“bib_date_year”）により、以下の 4 つの Excel ファイル（xlsb 形式）に分離されています。

WoS_NID_corres_1998_2004_ver.2020.1.xlsb：論文発表年が 1998～2004 年のデータ

WoS_NID_corres_2005_2010_ver.2020.1.xlsb：論文発表年が 2005～2010 年のデータ

WoS_NID_corres_2011_2015_ver.2020.1.xlsb：論文発表年が 2011～2015 年のデータ

WoS_NID_corres_2016_2019_ver.2020.1.xlsb：論文発表年が 2016～2019 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) 記事内アドレス番号(rs_address_seq)と記事内アドレス数(rs_address_count)：論文は共著であることが多く、また、一人の著者が複数の機関に所属することもあるので、一般に、同じ WoSCC 記事番号(WoS_ut)の下に複数の機関・アドレス情報のレコードが含まれます。rs_address_seq は、同じ記事中の機関・アドレスデータに WoSCC で付けられている一連番号、rs_address_count は、記事内の機関・アドレスデータの総数です。このテーブルでは、論文に外国機関に所属する著者が含まれる場合そのレコードが除かれ、日本所属機関のデータのみを収録していますので、rs_address_seq は飛び飛びになっていることがあります。このとき、rs_address_count は同一記事内のレコード数と一致しません。

なお、WoSCC には、著者所属を表すフィールド以外のフィールドに一部の著者所属データが含まれていることがありますが、このテーブルではそのデータは含んでいません。

(d) WoSCC の機関・アドレス情報：次の 3 つのフィールドがあります。

① 機関の名称(rs_organization)：代表機関の名称です。

② 組織の名称(rs_suborganization)：下部組織の名称です。

③ 機関のアドレス(rs_address)：機関とその所在地、郵便番号を示す情報です。rs_organization、rs_suborganization のデータも個々に含まれています。

なお、原則は上記の通りですが、rs_organization に機関と組織の両方の情報が入っていたり、rs_organization に組織名、rs_suborganization に機関名が入っていたりすることがあります。

(e) 同定フラグ：2.(3)で述べた L, M, R, V, S, H, N のいずれかです。L の場合 L1 と L2、M の場合 M1 と M2 がありますが、2 文字目の数字にはあまり意味がありません。同定フラグが S のレコードでは以下の(g), (h), (i)が、H または N のレコードでは以下の(f), (g), (h) , (i)が空白です。

(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。

(g) 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。

(h) 機関正式名:同定された機関の日本語正式名です。

(i) 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。

(j) 年内記事番号:出版年(bib_date_year)が同じ年の記事に、WoSCC 記事番号(WoS_ut)順に付けた一連番号です。

(k) 同定番号と同定数: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

[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。

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 と記事内機関番号は同じになります。

同定数は、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 99.2%で、残りが同定数 2～5 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(下記の表に同定フラグ H:65,476、同定フラグ N:60,257 を加えた 3,612,886)、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, M, R または V)は、全体の 95.0%です。また、機関同定されたうちでは、大学等 74.6%、公的機関 13.4%、その他の機関 12.0%となります。

セクター	同定フラグ					
	L	M	R	V	S	計
国立大学	1,721,654	94	7,492	134		1,771,661
国立短期大学	861	0				871
国立高等専門学校	10,332	58		5		10,506
公立大学	164,353	55	89	39		170,112
公立短期大学	896	2				926
公立高等専門学校	940	1		1		956
大学共同利用機関	50,755	0	75	12		51,451
私立大学	63,798	28	2	24	49	65,677
私立短期大学	390,752	132	5,392	89		406,053
私立高等専門学校	99,309	73	919	234	11,592	115,633
国の機関	597,506	414	1,436	73		618,446
国立研究開発法人等	2,749	0		1		2,780
地方自治体の機関	68	1				71
会社	195,827	0	873	48	44,978	246,795
非営利団体	106,293	267	4,591	73		115,179
その他の機関	1,712	2				1,742
計	3,407,805	1,127	20,869	733	56,619	3,487,153

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、クラリベイト社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、第一の場合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、1998-2019 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、1998-2019 の期間にわたる統計だけです。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) このテーブルのカバー率

このテーブルのデータは、2020 年 4 月時点における WoSCC カスタムデータから出版年 1998～2019 年のものを抽出しました。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、この期間についても、カバー率は 100%とはなっていません。2020 年 11 月時点の

WoSCC に含まれている日本論文（2(1)の条件(a)～(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。

出版年	WoSCC 日本論文数(2020.11 時点)	このテーブルに含まれる論文数	カバー率
1998	69,455	68,772	99.0%
1999	71,388	70,886	99.3%
2000	73,586	73,026	99.2%
2001	73,105	72,622	99.3%
2002	74,516	74,050	99.4%
2003	76,704	76,245	99.4%
2004	77,136	76,711	99.4%
2005	76,896	76,492	99.5%
2006	77,333	76,889	99.4%
2007	75,911	75,494	99.5%
2008	76,313	75,917	99.5%
2009	75,685	75,208	99.4%
2010	74,598	74,182	99.4%
2011	76,642	76,165	99.4%
2012	77,460	76,890	99.3%
2013	79,049	78,399	99.2%
2014	77,784	77,041	99.0%
2015	77,434	76,514	98.8%
2016	80,089	79,112	98.8%
2017	82,407	81,177	98.5%
2018	84,233	82,635	98.1%
2019	88,668	80,745	91.1%
計	1,696,392	1,675,172	98.7%

(2) 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、ごく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 2.0%でした。しかしその大部分は変遷前後の旧機関と新機関の間の同定違い（特に、東京工業大学の旧理学部と新理学院、旧工学部と新工学院など英語名が同一の新旧組織間の混同）で、代表機関が間違っていたものは 0.04%、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 0.003%でした。こ

のテーブルでは気付いたエラーを修正済みですので、エラーは恐らく 0.1%程度以下と思われますが、ご利用目的に応じてご自身でもデータの確認を行うようにしてください

今後も、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。

【WoSCC-NISTEP 大学・公的機関名辞書対応テーブル改訂履歴】

2017 年 4 月　WoSCC-NISTEP 大学・公的機関名辞書対応テーブル Ver.2017.1

2018 年 10 月　WoSCC-NISTEP 大学・公的機関名辞書対応テーブル Ver.2017.1.1

2019 年 12 月　WoSCC-NISTEP 大学・公的機関名辞書対応テーブル Ver.2017.1.2

2021 年 1 月　WoSCC-NISTEP 大学・公的機関名辞書対応テーブル Ver.2020.1