Scopus-NISTEP 大学・公的機関名辞書対応テーブル 説明書

2018 年 8 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”や “U”、“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 Scopus（Elsevier 社製）に含まれる機関名データから、国内の機関の同定（名寄せ）を行っています。この結果に基づいて、Scopus 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「Scopus-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。Scopus 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、エルゼビア・ジャパン株式会社の了解を得て、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:19,000 以上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。
- ○ 大学・公的機関名英語表記ゆれテーブル（Scopus 版）:約 200 の大学と約 50 の公的機関について、Scopus によく出現する機関名表記のゆれをまとめたデータです。

※このテーブルの利用について

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、Scopus 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 235 万件、所属機関データ数は延べ 467 万件です。

(a) 論文出版年が 1996～2016 年

(b) 日本の機関と判別されたもの（著者所属機関所属国が"jpn"）

(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びます)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（19,000 機関中約 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。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.(j)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。Scopus 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, E 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	Scopus 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。
E	3 の(f)に示す affiliation_1、affiliation_2、affiliation_3 がすべて空白のため、機関同定が不可能なデータ。

3. テーブルの構成

このテーブルは、論文の発表年(4. (b)の“year”)により、以下の6つのtsvファイルに分離されています。

Scopus_NID_corres_1996_2000.tsv: 論文発表年が 1996～2000 年のデータ
Scopus_NID_corres_2001_2004.tsv: 論文発表年が 2001～2004 年のデータ
Scopus_NID_corres_2005_2007.tsv: 論文発表年が 2005～2007 年のデータ
Scopus_NID_corres_2008_2010.tsv: 論文発表年が 2008～2010 年のデータ
Scopus_NID_corres_2011_2013.tsv: 論文発表年が 2011～2013 年のデータ
Scopus_NID_corres_2014_2016.tsv: 論文発表年が 2014～2016 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

このテーブルには、Scopus の著者所属機関データに含まれるデータ項目と、同定の結果 NISTEP で追加したデータ項目があります。それぞれの項目について説明します。

【Scopus の著者所属機関データに含まれる項目】

(a) scopus_eid:この著者所属機関レコードを含む Scopus 論文の論文識別番号です。
(b) year:Scopus 論文が原出版物に発表された年です。
(c) seq: 1 つの Scopus 論文(scopus_eid が同一)に含まれ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、1 つの scopus_eid に対しすべての順番が存在するとは限りません。
(d) Scopus_affiliation_id:Scopus によって判別された所属機関の識別番号です。
(e) address:所属機関の所在地(都道府県名、都市名など)です。
(f) affiliation_1、affiliation_2、affiliation_3:所属機関を示すデータです。affiliation_1 はすべてのレコードに存在しますが、affiliation_2、affiliation_3 が存在するレコードはそれぞれ 75%、30%程度です。3 つの項目にデータが存在する場合は affiliation_3 に、2 つの項目にデータが存在する場合は affiliation_2 に親機関名、それ以外に下部組織名が記されている場合が多いですが、必ずしもこれに従ってはいません。
(g) DOI:当該論文のデジタルオブジェクト識別子(Digital Object Identifier)です。約 80%の論文に付与されています。

【同定により追加された項目】

(h) 同定フラグ:2(3)で述べた同定レベルを示す記号で、L, M, S, H, N, E のいずれかです。同定フラグが S のレコードでは以下の(i), (k), (l), (m)が、H、N または E のレコードでは以下の(i), (j), (k), (l), (m)が空白です。
(i) 機関名辞書 ID(nid):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。
(j) セクター番号、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[1]。

[1] 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。

	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。

(k) 同定機関名：同定された機関の日本語正式名です。

(l) 代表/下位：同定された機関が 2.(2)で述べた代表機関の場合“TRUE”、下部組織の場合“FALSE”です。

(m) 代表機関名：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。

(n) 同定番号:一つの Scopus 所属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。分割されたレコードでは scopus_eid と seq は同じです。また、医学部と医学研究科のように、同じ代表機関の異なる下部組織が複数同定された場合も、同様の取扱いをしています。

(o) 同定数：上記の同定番号の繰り返し数です。

レコードは、第 1 ソートキー:year、第 2 ソートキー:scopus_eid、第 3 ソートキー:seq、第 4 ソートキー:同定番号により配列されています。

5. 同定結果の概要

論文出版年(year)別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 Scopus 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 92.0%です。

出版年	L	M	S	H	N	E	計	L+M比率
1996-1998	447,049	87	12,253	10,888	18,229	1,604	490,110	91.2%
1999-2001	476,089	427	11,795	10,660	20,114	1,746	520,831	91.5%
2002-2004	568,844	485	13,587	13,122	18,791	2,253	617,082	92.3%
2005-2007	717,257	493	17,258	15,132	25,335	2,953	778,428	92.2%
2008-2010	707,705	475	17,388	15,188	23,462	2,248	766,466	92.4%
2011-2014	763,695	333	17,362	18,053	22,076	1,580	823,099	92.8%
2014-2016	745,124	61	18,440	20,478	29,145	1,881	815,129	91.4%
計	4,425,763	2,361	108,083	103,521	157,152	14,265	4,811,145	92.0%

6. このテーブルの利用法

このテーブルには、主に次の 2 つの利用法が考えられます。

(1) Scopus 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。

第一は、Scopus 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。Scopus のカスタムデータを用いる場合は、検索したデータに scopus_eid の項目があります。これらの scopus_eid をこのテーブルの scopus_eid と接合することで、機関名の名寄せが可能となります。Scopus のオンラインデータを用いる場合は、検索した論文のうち分析したい論文にチェックをつけ、「エクスポート」を選択してください。「エクスポートする方法」から CSV を選ぶと、「エクスポートする情報」のなかに“EID”がありますので、これを含めてエクスポートすれば、scopus_eid が得られます。抽出された scopus_eid を、カスタムデータの場合と同様にこのテーブルと接合します。

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 scopus_eid の集合を作り、Scopus データベースからそれらに一致する scopus_eid のレコードを抽出します。これにより、Scopus 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。オンラインデータをご使用の方は、全データベースに対してこの方法を使うことはできませんが、まず広めに検索を行って、その結果をエクスポートすれば、その中の scopus_eid から所定のレコードを抽出することができます。

(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、1996-2016 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、セクター別の論文生産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、Scopus 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。一つの論文に同じ代表機関の異なる部局の著者が含まれている場合、この代表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。代表機関単位の論文数統計をとる場合には、

同じ scopus_eid の中の重複を削除する必要があります。

scopus_eid 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、1996-2016 の全期間にわたる統計だけです。期間、分野、論文の種類を区切った統計を得るには、Scopus 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 L または M)のエラー率は 0.3%未満です。

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ると幸いです。

(2) このテーブルのカバー率

このテーブルのもととなるデータは、2016 年末時点の Scopus カスタムデータです。したがって、出版年が 2017 年以降の論文については、このテーブルには含まれていません。また、Scopus では、適時データの追加、修正が行われていることから、2016 年以前についても、カバー率は 100%とはなっていません。

以下に 2018 年 8 月 14 日時点の Scopus に含まれている日本論文数(全分野の全ての論文タイプを対象)とこのテーブルがカバーする論文数の比較を示します。

2015 年までは 95%を超えていますが、2016 年のカバー率は約 91%、2017 年以降は 0%です。このようにこのテーブルのカバー率は年毎に変化しますので、このテーブルを利用して分析を行う際は、カバー率を必ず調べるようにしてください。

【参考】Scopus-NISTEP 大学・公的機関名辞書対応テーブルのカバー率

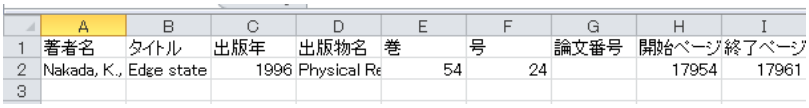
出版年	Scopus中の日本論文（2018年8月14日時点）	Scopus-NISTEP大学・公的機関名辞書対応テーブル（2016年末時点）	カバー率
1996	87,475	85,477	97.7%
1997	93,785	91,434	97.5%
1998	94,075	90,933	96.7%
1999	97,246	93,669	96.3%
2000	100,790	96,977	96.2%
2001	99,625	95,505	95.9%
2002	103,033	97,787	94.9%
2003	108,998	104,985	96.3%
2004	112,308	108,632	96.7%
2005	123,377	120,955	98.0%
2006	124,939	123,021	98.5%
2007	121,170	119,200	98.4%
2008	120,836	119,188	98.6%
2009	126,049	124,348	98.7%
2010	127,397	125,855	98.8%
2011	130,301	128,434	98.6%
2012	131,689	129,952	98.7%
2013	133,897	131,386	98.1%
2014	130,404	129,220	99.1%
2015	125,217	121,121	96.7%
2016	127,671	116,067	90.9%
2017	126,010	0	0.0%

8. scopus_eid の取得方法について

- scopus_eid は、Scopus に収録されている論文についてのユニークな ID 番号です。Scopus の検索画面(http://www.scopus.com/home.url)から、次の方法で取得可能です。なお、Scopus を利用するには、エルゼビア・ジャパン社との契約が必要です。
- Scopus の検索で得られた論文の内、書誌情報等をダウンロードしたい論文にチェックをつけ、「エクスポート」を選択してください。

- 次にエクスポートの形式と出力内容を選んでください。この時、書誌情報を出力内容に必ず含めてください。「エクスポート」ボタンを押下するとデータが出力されます。

- ダウンロードした書誌情報には、Scopus 上の該当論文の URL の情報が含まれています。この URL の中で、下線を引いた部分が scopus_eid に該当します。形式として、コンマ区切りファイルを選ぶことで、エクセルで編集可能なデータがダウンロードできます。



	A	B	C	D	E	F	G	H	I
1	著者名	タイトル	出版年	出版物名	巻	号	論文番号	開始ページ	終了ページ
2	Nakada, K.,	Edge state	1996	Physical Re	54	24		17954	17961
3									

J	K	L	M	N	O	P	Q	R
ページ数	被引用数	DOI	リンク	文献タイプ	アクセスター	情報源	EID	
	2833	10.1103/PI	https://ww	Article		Scopus	2-s2.0-0000781318	