

大学・公的機関名英語表記ゆれテーブル(Ver.2015.1)利用の手引き

2015年10月15日

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。この問題に対処するため、Web of Science Core Collection (WoSCC) や Scopus などのデータベースでは、各機関に固有の識別番号を与える、この識別番号による所属機関の一括検索機能を設ける等、機関検索の精度向上が図られています。しかし、あるテーマに関する検索結果をダウンロードしてその中の機関構成を分析したいときなどは、これらの機能を使うことができません。また、機関の下部組織等についてまで識別がなされているわけではありません。

科学技術・学術政策研究所(NISTEP)が「科学技術イノベーションにおける“政策のための科学”推進事業」の一環として進めている「公的研究機関に関するデータ整備」では、その整備事業の中で、データベースで使われている機関表記のゆれを調査・分析しています。これまで、Scopus データベースと WoSCC データベースに現れる機関名英語表記ゆれを調査した結果の一部を、それぞれ Elsevier 社、トムソン・ロイター社の了解を得て「大学・公的機関名英語表記ゆれテーブル(Scopus 版)」及び「大学・公的機関名英語表記ゆれテーブル(WoSICC 版)」¹として公開してまいりました。また、このサイトで公開されていた「NISTEP 大学・公的機関名辞書(Ver.2014.1)」にも、NISTEP が種々の情報源から調査した別名(機関の通称、略称等)のデータが含まれていました。

今回、これらの別名データと大学・公的機関名英語表記ゆれテーブル(WoSICC 版)を統合して、「大学・公的機関名英語表記ゆれテーブル(Ver.2015.1)」(以下、混同の恐れがなければ単に「テーブル」と略記)として公開することといたしました。これに伴い、NISTEP 大学・公的機関名辞書の更新版(Ver.2015.1)では別名データは除かれています。

なお、「大学・公的機関名英語表記ゆれテーブル(Scopus 版)」については、当面統合せずそのまま公開しております。これは、Scopus 版が大学と公的機関のみを対象にしていること、機関の表記ゆれの中に一部その下部組織に対する表記が含まれていることなど、データ収録の基準にやや異なる点があるためです。これらの点を修正した後、すべてを統合した機関名英語表記ゆれテーブルとする予定です。

¹ 公開しているテーブルでは「WoSCC 版」ではなく「Web of Science 版」となっています。公開当時のデータベース名は Web of Science でしたが、その後 Web of Science Core Collection と改称されました。混乱を避けるためここではすべて「WoSCC 版」と記述します。

※このテーブルの利用について

大学・公的機関名英語表記ゆれテーブル(Ver.2015.1)の利用については、クリエイティブ・コモンズ・ライセンス(CC ライセンス)の「表示-継承」を適用します。すなわち、原作者のクレジット



を表示し、改変した場合にはこのテーブルと元の作品と同じ CC ライセンス (表示-継承) で公開することを守れば、営利目的での二次利用も許可されます。表示するクレジットは次のようになります。

原作者名: 文部科学省科学技術・学術政策研究所

作品タイトル: 大学・公的機関名英語表記ゆれテーブル (Ver.2015.1)

URL: <http://www.nistep.go.jp/research/scisip/data-and-information-infrastructure>

CC ライセンスと、このライセンスのコモンズ証、リーガルコードについては、

<http://creativecommons.jp/licenses/> をご覧下さい。

2. 収録対象機関

研究活動を行っている我が国の機関が対象です。大学等(短期大学、高等専門学校、大学共同利用機関を含む)、公的機関(国の機関、特殊法人・独立行政法人等)の他、地方自治体の機関、非営利法人、民間企業²も含まれます。以下では、これらの独立した機関を「代表機関」、代表機関に属する組織を「下部組織」と呼びます。単に「機関」というときは、代表機関と下部組織の両方を意味します。このテーブルには、大学、大学共同利用機関、国の機関、特殊法人・独立行政法人の主要な下部組織も含まれています。

NISTEP 大学・公的機関名辞書(以下「機関名辞書」と略記)の区分に従い、以下の機関は下部組織ではなく代表機関としています。

- (a) 大学の一部としての短期大学部、高等専門学校
- (b) 国立高等専門学校: これらは独立行政法人国立高等専門学校機構の下部組織ですが、ここでは代表機関としています。
- (c) 国立試験・研究機関: 但し、試験・研究機関に属しない国の機関(気象庁地磁気観測所等)は属する省庁の下部組織としています。
- (d) 地方公共団体の公設試験研究機関等

3. データ源と抽出した機関表記データ

1. で述べたように、このテーブルは、もともと機関名辞書に含まれていた機関表記(テーブルのデータ源が「辞書」となっているデータ)と WoSCC 版に含まれていた機関表記(テーブルのデータ源が「WoS」となっているデータ)を合体したものです。ここでは、それぞれのデータ源における機関表記データについて説明します。

3.1 機関名辞書の機関表記データ

機関名辞書に含まれている英語の正式機関名称とそれ以外の「別名」(通称、略称等)です。正

² 民間企業については、このサイトの「NISTEP 企業名辞書」で詳しいデータが公開されていますので、NISTEP 大学・公的機関名辞書(Ver.2015.1)には含まれていません。しかし、WoSCC データベース等の利用では企業の英語名データも重要なので、このテーブルには含めています。

式名称の場合、テーブルの「正式名称」欄に○印が付いています。但し、○印が付いているのは当該機関の Web サイトにより正式名称であると確認された表記なので、○印が付いていなくても正式名称である場合があります。地方公共団体の機関、会社、非営利法人に属する機関の表記にはこのような例が多く含まれます。

別名はいろいろな情報源から採られています。NISTEP では、機関名辞書を用いてデータベース中の著者所属機関データの同定を行っていますので、その際同定漏れや誤同定を防止するためにいろいろな別名データを収録しています。

正式名称か別名かに拘わらず、会社を表す“Co., Ltd.”、“Corp”等を除いて、機関の法人格を示す接頭辞または接尾辞は原則として付いていません。大学の下部組織名称には、下部組織名の後に上位の大学名が付けられていますが、その他の機関では、下部組織名称に代表機関名が含まれているかどうかは不定です。

[例] Graduate School of Medical Sciences, Kyushu University

3.2 WoSCC 版の機関表記データ

データ源は、1996-2012 年の期間に WoSCC データベースに採録された論文のうち、日本の機関に所属する著者を含む論文のデータです。これに該当する論文は約 150 万件、その中の日本機関のデータは延べ 275 万件以上になります³。

これらの論文における著者所属機関データの機関同定(所属機関表記を機関名辞書の登録機関に対応付けること)によって得られた結果が、このテーブルに含まれるデータです。但し、同定される表記データの種類は膨大な数になるので、このテーブルに記載する機関表記を次の手順で選択しました。

- (1) 同定された機関データが延べ1,000以上の機関を選択します。このとき、下部組織はその代表機関に、現存しない機関は最終継承機関に合体して集計します。
- (2) (1)で選択した機関データから、10回以上出現した表記バリエーションを取り出します⁴。
- (3) これらの表記バリエーションを、機関名辞書に登録した機関単位に分けます。代表機関とその下部組織、非現存機関とその継承機関は、それぞれ別の機関単位です。

これにより選択された表記バリエーションは2,217となりますが、それらによりカバーされる WoSCC 中の出現頻度は、日本機関データ全体の68%、このテーブルに存在する機関のデータの75%になります⁵。つまり、WoSCC でこれらの表記バリエーションを含めた検索や抽出を行えば、対象の機関の大部分の論文データをカバーすることができます。

³ WoSCC の書誌データの利用については、利用機関とトムソン・ロイターが取り交わす利用規約が適用されます。

⁴ 但し、下部組織については、代表機関の直下の下部組織の表記のみを取り出しています。たとえば、大学院研究科名を表す Grad Sch Sci という表記バリエーションは取り出しますが、研究科名とその下の専攻科名が結合した Grad Sch Sci Dept Phys や、専攻科名のみ Dept Phys という表記バリエーションは原則として取り出していません。

⁵ 代表機関のカバー率はこれよりずっと高いのですが、下部組織では表記は極めて多様なためカバー率が低くなります。下部組織の表記ゆれの分析については、『NISTEP 大学・公的機関名辞書の整備とその活用—大学下部組織レベルの研究データ分析に向けて—』, NISTEP NOTE-15, 科学技術・学術政策研究所, 2015年10月 (<http://www.nistep.go.jp/wp/wp-content/uploads/NISTEP-NN15-FullJ.pdf>)を参照してください。

4. テーブルの概要

4.1 機関数と表記バリエーション数

機関名辞書の機関表記と WoSCC の機関表記を合体する際、以下の処理を行いました。

- (1) 二つのデータ源に全く同じ表記がある場合は、その重複を除きました。
- (2) このテーブルに含まれる機関は、機関名辞書の別名あるいは WoSCC 表記データを含む機関に限っています。つまり、正式名称以外の表記データがない機関は含んでいません。但し、正式名称しかない機関が代表機関であって、別名あるいは WoSCC 表記データを含む下部組織がある場合は、その代表機関もテーブルに含まれます。このテーブルにない機関の英語正式名称は、このサイトから公開している機関名辞書によって知ることができます。

テーブルに記載される機関は 10,022 (代表機関 8,934、下部組織 1,088) です。また、表記バリエーションの総数は 14,563 (代表機関 11,813、下部組織 2,750) で、12,431 が機関名辞書から、2,217 が WoSCC からのデータです (両者に含まれている表記が 85)。

4.2 データ項目

テーブルには以下のデータ項目が含まれます (それぞれの内容については 5. をお読み下さい)。

[A] 機関表記に関するデータ項目

- ・表記バリエーション
- ・正式名称であるかどうか
- ・データ源が機関名辞書、WoSCC のいずれか
- ・WoSCC における出現頻度 (データ源が WoSCC の場合)

[B] 機関に関するデータ項目

- ・機関 ID
- ・機関の日本語正式名称
- ・機関が属するセクター
- ・代表機関か下部組織かの別 (下部組織の場合その代表機関を記載)
- ・現存機関か非現存機関かの別 (非現存機関の場合その最終継承機関を記載)

5. テーブルの見方

テーブルは Excel 形式 (.xlsx) で単独のシートから成ります。データ項目 (テーブルの各列) は、機関表記に関する項目と機関に関する項目に大別されます。

5.1 各データ項目の説明

[A] 機関表記に関するデータ項目

- (1) 表記バリエーション: 機関名辞書または WoSCC から抽出された機関表記です。詳しくは 3. をお読み下さい。
- (2) 機関内番号: 同一機関の表記バリエーションに付けられる一連番号です。正式名称がある場合はその番号が 1、あとはアルファベット順です。
- (3) 正式名: その表記バリエーションが正式名称の場合 “○” が付いています。正式名称については 3.1 の説明をお読み下さい。

- (4) データ源:その表記バリエーションが採られたデータ源で、次のいずれかです。
- ・辞書、WoS:機関名辞書とWoSCCの両方に含まれる表記バリエーション
 - ・辞書:機関名辞書のみに含まれる表記バリエーション
 - ・WoS:WoSCCのみに含まれる表記バリエーション
- (5) WoS 頻度:データ源が「辞書、WoS」または「WoS」の場合、その表記バリエーションがWoSCCに出現した回数です。詳細は3.2の説明をお読み下さい。

[B] 機関に関するデータ項目

以下の項目についての詳細な説明は、このサイトにある「NISTEP 大学・公的機関名辞書 (Version 2015.1)利用マニュアル」をご覧ください。

- (1) 機関ID:機関名辞書で与えられている各機関の識別番号です。番号に特別の意味はありません。
- (2) 機関名:機関(代表機関または下部組織)の日本語正式名称です⁶。
- (3) セクター番号及びセクター:機関が属するセクターとそれらに付けられた番号で、次のようになっています。

	セクター番号	セクター
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	特殊法人・独立行政法人
その他の機関	10	地方自治体の機関
	15	会社
	16	非営利団体
	17	その他の機関

- (4) 代表機関:“○”が付いている機関は代表機関です。下部組織の場合は空白です。
- (5) 代表機関名:下部組織の場合、それに対する代表機関の名称です。代表機関の場合は「機関名」の名称と同じです。
- (6) 現状:機関が現存する場合は空白、現存しない場合は“No”です。

⁶ 2015年4月1日にいくつかの独立行政法人が国立研究開発法人に移行しましたが、このテーブルの機関名は、機関名辞書の現公開バージョン(Version 2015.1)に基づいており、それは2015年1~2月時点の調査によるものであるため、すべて独立行政法人の名称になっています。

- (7) 最終継承機関:「現状」が“No”の場合、その継承機関があれば、最終の継承機関を記入しています。

5.2 レコード(テーブルの行)の単位と並び順

一つのレコード(テーブルの 1 行)は一つの表記バリエーションに対応します。複数の表記バリエーションを持つ機関はその数だけの行数を占め、機関内番号で各行が識別されます。このような機関では、6.1 の[B]に示した機関に関するデータ項目には同じデータが並ぶことになります。やや冗長ですが、自由にソートができるようこのような構成にしています。

レコードは次の順序に配列されています。

- ① セクター番号の順
- ② 各セクターの中で代表機関名の漢字コード順
- ③ 各代表機関の中でまず代表機関(○印)、残りは機関(下部組織)名の漢字コード順
- ④ 各機関内でまず英語正式名称表記(○印)、残りは表記バリエーションのアルファベット順

【大学・公的機関名英語表記ゆれテーブル改訂履歴】

2012/12/18 NISTEP 大学・公的機関名辞書 Ver.2012.1

2013/12/10 大学・公的機関名英語表記ゆれテーブル(WoSCC 版)(Ver.2013.1)

2014/11/14 NISTEP 大学・公的機関名辞書 Ver.2014.1

2014/11/14 大学・公的機関名英語表記ゆれテーブル(WoSCC 版)(Ver.2014.1)

2015/10/15 NISTEP 大学・公的機関名辞書 Ver.2014.1 の英語名称データと大学・公的機関名英語表記ゆれテーブル(WoSCC 版)(Ver.2014.1)を統合し、大学・公的機関名英語表記ゆれテーブル(Ver.2015.1)として公開

【付録】このテーブルのデータを WoSCC の検索に利用する場合の注意

WoSCC の機関検索にこのテーブルに含まれる表記バリエーションを利用すれば、かなり高い再現率を得ることができます。ここでは、その際に注意すべき点について述べます。

A1. 機関名の略記について

WoSCC では、機関名の表記に次のような略記法を用いています。このテーブルにおける WoSCC をデータ源とする表記もこの略記法に従っています(但しそれからゆれた表記もあります)

(1) よく出現する語の短縮形表記

代表的な例は次の通りです。

University → Univ

Institute → Inst

National → Natl

Science, Scientific → Sci

Technology, Technological → Technol

Medicine, Medical, Medicinal → Med

And → &

WoSCC の検索で検索項目に所属機関を選ぶと、「短縮形リストを表示」という案内が示されます。このリストから、使用されている短縮形を知ることができます。但し、検索をする場合は原形、短縮形のどちらを用いても構いません(検索語に University、Univ のどちらを用いても同じ結果が得られます)。

(2) 前置詞、冠詞の省略

前置詞の of、for、冠詞の the 等は省略されます。

(3) ハイフン、アポストロフィー等の省略

語中に含まれるハイフン(-)やアポストロフィー(')等は省略されます。ハイフンを省略したとき、その両側を詰める場合とスペースに置き換える場合があります。

このような略記を用いているため、機関名はたとえば次のように表示されます。

The University of Tokyo → Univ Tokyo

National Institute of Advanced Industrial Science and Technology

→ Natl Inst Adv Ind Sci & Technol

The University of Electro-Communications → Univ Electrocommun

Nara Women's University → Nara Womens Univ

A2. 代表機関と下部組織の表記

(1) WoSCC の著者所属機関データフィールド

WoSCC の XML データファイルでは、著者所属機関データを、affiliation_org、affiliation_suborg などいくつかのサブフィールドに分割しています。原則として、affiliation_org サブフィールドには代表機関、affiliation_suborg サブフィールドには下部組織の名称が記載されていますので、このテーブルの代表機関と下部組織の表記も、多くはそれに従って抽出されています。この原則に従っていない場合(affiliation_org サブフィールドと affiliation_suborg サブ

フィールドの間で下部組織名と代表機関名が逆転している場合、affiliation_suborg サブフィールドに代表機関と下部組織の合体した名称が入っている場合などは、修正して記載しました。

(2) affiliation_org サブフィールドの表記

次のように、代表機関以外を示す語が混入している表記も含まれています。

- 機関名と下部組織名が合体した表記

[例] Hirosaki Univ Hosp
Kagoshima Univ Museum
Kinki Univ Sch Med

- 所在地の一部が混入した表記

[例] Kyoto Univ Katsura
JAERI Takasaki

(3) affiliation_suborg サブフィールドの記述

原則として、その代表機関の情報は含まれていません。例えばある大学の工学研究科の場合は Grad Sch Engn となっています。これだけで検索すればあらゆる大学の工学研究科のデータが出てくることになります。まず、目的とする大学の表記を代表機関テーブルから選択して検索した後、その結果に対して下部組織の表記で絞り込み検索を行うなどの方法をとります。

大学の affiliation_suborg サブフィールド表記には、大学直下の組織(学部や大学院研究科)ではなく、その下の学科や専攻科が記載されていることが多いのですが、このテーブルに含まれるのは、原則として代表機関直下の下部組織です(例外もあります)。

A3. 混同しやすい英語機関名

次のように、異なる機関が同一または類似の英語名称を持つと、それらの名称で検索した場合ノイズ(目的以外の機関の混入)や検索洩れが生じやすくなりますので、注意が必要です。

(1) 同一名称を持つ異なる機関

統合や改組を行った機関が、日本語機関名は変更したのに英語名はそのままという例はよくあります。たとえば、東京都立大学と首都大学東京はどちらも Tokyo Metropolitan University です。また、国立研究所の宇宙科学研究所は独立行政法人宇宙航空研究開発機構の下部組織に移行しましたが、Institute of Space and Astronautical Science の英語名はそのままです。

このような継承関係がない機関が全く同じ英語名を持つことは、大学や公的機関ではほとんどありませんが、会社や非営利団体ではいくつかあります。

(2) 類似の名称を持つ機関

最も注意を要する例として静岡大学と静岡県立大学があります。英語の正式名称はそれぞれ Shizuoka University、University of Shizuoka なので、WoSCC ではそれぞれ Shizuoka Univ、Univ Shizuoka と表記されることとなりますが、これらに所属する著者が、論文の所属機関に誤って違う大学の名称を記載していることがあります。このような誤記は、データベース作成時にも修正が困難ですので、この両大学では、一方の名称で検索した結果にはある程度のノイズや検索洩れがあります。

他にも、東京農工大学(WoSCC の表記は Tokyo Univ Agr & Technol)と東京農業大学(WoSCC の表記は Tokyo Univ Agr)なども混同しやすい例です。

(3) 英語名がごく一般的な機関

たとえば、分子科学研究所、厚生労働省国立公衆衛生院（現在は厚生労働省国立保健医療科学院）の WoSCC での表記は、それぞれ、Inst Mol Sci、Inst Publ Hlth です。しかし、これらに含まれる語はいずれも機関名によく使われるため、類似した名称を持つ機関が多く、このまま検索すると大量のノイズを生ずるので注意が必要です。