概要

（裏白紙）

概要

科学技術・学術政策研究所（NISTEP）が、文部科学省の「科学技術イノベーション政策における『政策のための科学』推進事業（SciREX）」の一環として 2011 年度以来進めている「大学・公的機関における研究開発に関するデータ整備」は、NISTEP 大学・公的機関名辞書（以降「機関名辞書」、あるいは誤解ない場合単に「辞書」という）の整備、この辞書を用いた NISTEP 機関同定プログラム（以降「名寄せプログラム」あるいは単に「プログラム」という）の開発、及びそれらの活用の普及を中核的業務としている。「大学・公的機関における研究開発に関するデータ整備」で得られた機関名辞書等の各種データは「https://www.nistep.go.jp/research/scisip/randd-on-university」にて公開されている。機関名辞書の 2020 年 6 月（ver.2020.1 の公開月）から 2022 年 11 月までの月平均ダウンロード数は 43.9 回である。多くはダウンロードした各機関の内部で利用されていると見られるが、NISTEP で把握している利用例としては、内閣府で開発している e-CSTI での利用、雑誌やデータ等の分析における著者所属機関の分類での利用、学術研究での利用などがある。また、名寄せプログラムの利用登録者数は、大学の研究アドミニストレーター(URA)を中心に 30 名以上（2022 年 1 月時点）である。

研究開発の動向を把握するため、種々の情報源を用いて機関レベル、組織レベルにデータを整理・分析しようとすると、機関名のゆらぎ、下部組織情報の不足、機関の変遷の把握の困難さ、セクター情報の不足などの問題点に直面する。これらの問題点があるため、情報源に記述される機関名を用いて特定の機関のデータを抽出しようとしても、目標の機関の洩れ、目標以外の機関の混入が起こる。機関名辞書と名寄せプログラムは、様々な方法を用いて精確な機関同定を行う手段を提供する。

例えば、データ分析でよく用いられる Scopus データベースを用いて、東京大学医学系研究科所属著者の論文を得たい場合、その名称である"Graduate School of Medicine, the University of Tokyo"で検索しただけでは多くの洩れが生ずる。その主な理由は、Scopus ではこの研究科名を省いてその下の専攻名や教室名（" Department of Surgery "等）を表示していることがあるためであるが、機関名辞書と名寄せプログラムではこの点を考慮している。Scopus の 2020-2021 年発表論文データでは、名寄せプログラムで東京大学医学系研究科に同定された 4,822 件の 2 割弱に及ぶ 928 件がこの種の表記であった。

もう一つの例として、青森県六ヶ所村にある公益財団法人環境科学技術研究所を挙げる。この研究所の英語名は"Institute for Environmental Sciences"であるが、この名称は機関名によく使われる単語のみから成るので、他の機関と混同されやすい。そのため、機関名辞書と名寄せプログラムでは類似の名の機関と識別するための処置を行っている。やはり Scopus の 2020-2021 年発表論文データにおいて単純に上記英語名で検索すると 142 件がマッチするが、名寄せプログラムを用いればそのうち 51 件のみがこの研究所に同定され、他の 71 件はそれ以外の機関に該当することが判る。

NISTEP では 3 つの目標の下に機関名辞書と名寄せプログラムの整備を進めている。第一は、研究開発を行う国内の主要な機関についての基本的情報を系統的・継続的に取得し、アーカイブ化することである。第二は、我が国の科学技術政策立案・検討の基礎資料として活用されることである。そして第三

は、我が国の研究開発推進の基盤として広く利用されることである。

このレポートは、主にこの第三の目標に関係する。機関名辞書は、2012 年にリリースして以来毎年 1～2 回公開データを更新している。一方名寄せプログラムは、主に Web of Science Core Collection（以降 WoSCC と略す）及び Scopus データベースの所属機関データの名寄せに NISTEP 内部で利用し、その成果データを公開しているが、2021 年度からプログラム自身の公開を開始した。これらの利用者及び潜在利用者にこのレポートを活用していただき、利用の促進に繋がれば幸いである[1]。

1　機関名辞書の概要

1-1　収録対象とする機関

機関名辞書の収録対象は、研究開発を行っている国内に所在する機関である。「大学・公的機関名辞書」という名のように、大学等（短大、高専、大学共同利用機関を含む）と公的機関（国の機関及び国立研究開発法人等（独立行政法人、特殊法人を含む）を指す）に主力を置くが、研究開発を行う地方公共団体の機関、民間企業、非営利法人等もできるだけ収録する。

機関名辞書の収録機関については次の 2 つの特徴を持つ。

(1) 独立した機関（「代表機関」という）のほか、その下部組織も収録の対象とし、上位機関との関係を付ける。特に、主要な大学、大学共同利用機関、国立研究開発法人、病院機構の下部組織は包括的に収録する。以下では、単に「機関」と言えば代表機関、下部組織を合わせて意味するものとする。

(2) 統廃合や名称変更があって非現存となった機関も保持し、継承の機関がある場合はそれと関係づけをする。この種の情報は、機関の活動を時系列で追跡するときに有用である。

2022 年 6 月時点における機関名辞書の収録機関数を概要図表1に示す。

1-2　主な収録情報とその特徴

1-2-1　セクターと病院フラグ

各機関を概要図表 1 に示す 17 のセクターのいずれかに分類する。これとは別に病院フラグを設け、病院である機関はその値を"True"、それ以外の機関は"False"とする。

1-2-2　機関の日本語名称

各機関には必ず 1 個の日本語正式名が与えられる。この正式名は当該機関が名乗っている名称によるが、法人格付与の有無、下部組織の名称（必ず先頭に代表機関名を付ける）等で統一を行っている。

---

[1] プログラム公開に先だって行った試用実験の終了時に行ったアンケート調査において、名寄せにおいて注意すべき点や名寄せプログラムの性能等について知りたいという意見があったことが、この報告書作成のきっかけとなった。ご意見をくださった方に謝意を表する。

概要図表１　セクター別、代表・下位別、現存・非現存別収録機関数（2022 年 6 月現在）

セクター	代表機関			下部組織			合計		
	現存	非現存	小計	現存	非現存	小計	現存	非現存	総計
1 国立大学	86	15	101	1,567	640	2,207	1,653	655	2,308
2 国立短大		26	26					26	26
3 国立高専	51	8	59				51	8	59
4 公立大学	99	19	118	88	13	101	187	32	219
5 公立短大	15	49	64				15	49	64
6 公立高専	3	4	7				3	4	7
7 大学共同利用機関	4	3	7	28	1	29	32	4	36
8 国の機関	54	72	126	64	17	81	118	89	207
9 国立研究開発法人等	80	86	166	396	202	598	476	288	764
10 地方公共団体の機関	792	261	1,053	314	109	423	1,106	370	1,476
11 学校法人	672	27	699				672	27	699
12 私立大学	633	81	714	523	126	649	1,156	207	1,363
13 私立短大	299	286	585				299	286	585
14 私立高専	3	1	4				3	1	4
15 会社	4,078	1,009	5,087	14	5	19	4,092	1,014	5,106
16 非営利団体	3,701	3,744	7,445	91	54	145	3,792	3,798	7,590
17 その他	9	2	11	1	1	2	10	3	13
計	10,579	5,693	16,272	3,086	1,168	4,254	13,665	6,861	20,526

1-2-3　機関の英語名称

名寄せに重要な役割を果たす英語名称を次の 4 種に区分し、その充実に努めている。

(1)　正式名(Formal)：当該機関が定める英語名。Web サイト等から確認できた場合のみ正式名とする。

(2)　別名(Alias)：正式名と思われるが未確認の名称、正式名ではないがよく使われる通称、旧名、略称、正式名称中の一部の語を置き換えまたは省略した名称など。

(3)　揺らぎ名(Variant)：(1)、(2)以外に、名寄せに用いるためのいろいろな名称で、正しい表記ではないものも含む。

(4)　非使用名(NotUse)：機関名辞書に収録するが、他の機関との混同の恐れがあるため名寄せには使用しない名称。

大学の下部組織の英語名には、原則として(Formal の場合は必ず)末尾に大学名を付ける。これは、大学下部組織の同定を確実にするためである。

1-2-4　機関の階層関係を表す情報

ある機関に対して、その直上及び直下の機関との関係づけを行う。代表機関の階層を"1"、その直下の機関の階層を"2"、第 2 階層機関の直下の機関の階層を"3"とする。代表機関に対しては「代表機関フラグ」の値を"True"、下部組織に対しては"False"とする。

1-2-5　機関の変遷情報

ある日本語正式名を持つ機関が廃止または統合され、あるいはその名称が変更されることにより存在しなくなったとき、機関の変遷が起こったとする。現存していない機関には下記の情報項目を付記する。

(1) 移行区分:「統合」、「廃止」、「変更」のいずれか。

(2) 移行年月日:移行が発生した年月日を YYYY-MM-DD の形式で記載する。

(3) 継承機関:移行区分が「統合」と「変更」の場合は必ず継承機関が存在する。「廃止」の場合も、その事業等を実質的に引き継いだ機関があればそれを継承機関とする。

1-2-6　その他の情報

(1) 機関の所在地の郵便番号

(2) 大学の下部組織種別:大学の下部組織には以下のいずれかを付与する:①学部;②大学院;③専攻科・別科;④学部・大学院統合;⑤教員組織;⑥研究所;⑦全学組織; ⑧病院。共同利用・共同研究拠点または世界トップレベル研究拠点形成プログラム(WPI)に指定された組織には、上記下部組織種別の後に括弧書きで「拠点」と付記する。

(3) 所管・被所管関係:国立試験研究機関(あるいは独立行政法人)とそれを所管する省庁、県の公設研究機関と県庁の間の関係づけである。この関係づけがなされた機関が同時同定されると、所管される方の機関にのみ同定がなされる。

(4) 外部の機関識別情報:NISTEP 企業名辞書の企業 ID、JISX0408(大学・高等専門学校コード)、科研費機関番号の 3 つの外部情報源の機関 ID と対応付けを行っている。

1-3　代表機関と下部組織

1-3-1　組織形態としては下部組織であるが代表機関扱いとする機関

以下の機関は、機関名辞書では下部組織ではなく代表機関とする。

- 大学の一部としての短期大学部または高等専門学校
- 国立高等専門学校
- 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は、属する省庁の下部組織とする。
- 地方公共団体の公設試験研究機関、公立病院等

1-3-2　下部組織収録の考え方

研究開発を行っている主要な下部組織を収録する。以下の組織は必ず収録する。

(1) 大学の下部組織

大学下部組織のうち、①附属病院、②国立大学の附置研究所、③共同利用・共同研究拠点及び WPI 拠点に指定された組織は、階層の如何に関わらず収録する。

下記の 32 大学については、すべての第 2 階層組織(事務組織あるいは統括・管理組織と判断されるものは除く)を収録する。

［国立］北海道大学、東北大学、筑波大学、群馬大学、千葉大学、東京大学、東京医科歯科大学、東京工業大学、東京農工大学、新潟大学、富山大学、金沢大学、福井大学、信州大学、岐阜大学、名古屋大学、京都大学、大阪大学、神戸大学、岡山大学、広島大学、徳島大学、九州大学、長崎大学、熊本大学

［公立］大阪公立大学（前身の大阪市立大学、大阪府立大学を含む）

［私立］慶應義塾大学、早稲田大学、東海大学、東京理科大学、日本大学、近畿大学

これらは発表論文数の多い大学を選んだものであるが、福井大学は、下部組織更新情報の提供において協力を得られたことによる。

(2) 大学共同利用機関、国の機関、国立研究開発法人等の下部組織

以下の組織（ほとんど第 2 階層）を収録する。

- 大学共同利用機関である 4 つの機構の研究所等
- 大規模な国立研究開発法人（前身の独立行政法人を含む）の主要な組織
- 国の機関及び独立行政法人（認可法人を含む）に所属する病院及び大学校

1-4　情報収集、公開、ファイル形式

1-4-1　機関名辞書更新のための情報収集

(1) 定期的な一斉調査による更新

大学、短大、高専、大学共同利用機関、学校法人については毎年 10～11 月、国の機関、国立研究開発法人等については、毎年 1 月に Web 調査を行い、代表機関、下部組織の新設と廃止に関する情報、英語名、郵便番号等の変更の情報を収集する。会社については、NISTEP 企業名辞書との連携により、毎年変更情報を入手する。それ以外のセクター（地方公共団体の機関、非営利団体、その他）の機関についての一斉調査はそれぞれ 3～4 年に 1 回に留めている。

(2) 名寄せ結果等からのデータ補充

毎年度はじめに WoSCC と Scopus の著者所属機関の名寄せを行い、そのチェックの結果を機関名辞書にフィードバックしている（4 を参照）。その他、種々の調査中あるいは日常業務中に気づいた修正情報を記録しておき、適時更新を行う。

1-4-2　機関名辞書の公開とファイル形式

2012 年 12 月に初版を公開して以降、年に 1、2 回更新版を公開している。現在の版は 2022 年 6 月に公開した。また、2021 年 1 月に初めての英語版を公開し、2022 年 6 月に更新した。

機関名辞書の元ファイルは関係データベース型の 6 ファイル（tsv 形式）から成るが、このままのでは一覧性に欠けるため、Excel の 1 シートに加工して公開している。名寄せプログラムの利用者には、元ファイルのフルデータを提供している。

2　NISTEP で実施している名寄せの方法

NISTEP では、自ら開発する名寄せプログラム（国内の研究機関の英語表記データが対象）により WoSCC 及び Scopus の著者所属機関データの名寄せを実施してその結果を公開するとともに、2021 年 12 月からプログラム自身も公開している。利用者にはプログラム公開サイトへのログイン ID を発行し、ログイン後プログラムや辞書をダウンロードして利用してもらう方法を採っている。

このプログラムの特徴は以下の通りである。

- WoSCC と Scopus のデータに傾注してプログラム開発を行ってきたが、国内機関の英語表記データであれば一般に適用可能である。
- 機関名辞書に登録されている下部組織や非現存機関も同定対象となり、これは他の名寄せプログラムにはあまり見られない特徴である。
- 第 1 種の過誤（誤同定）と第 2 種の過誤（同定洩れ）はトレードオフの関係にあるが、第 1 種の過誤を避ける方を優先している。

2-1　著者所属機関データのフィールド構成

名寄せに用いる WoSCC と Scopus のファイルは、それぞれクラリベイト・アナリティクス・ジャパン株式会社、エルゼビア・ジャパン株式会社から購入した XML 形式のデータファイルを、tsv ファイルに変換したものである。名寄せで重要な著者所属機関を示すデータはいくつかのフィールドに分割されている。

まず、所属国フィールドを用いて日本の機関のみを抽出し、名寄せを行う。WoSCC の場合、名寄せで主に用いるのは、代表機関の名称を示すフィールド（以下 ORG という）、下部組織の名称を示すフィールド（以下 SUBORG という）、全アドレス情報を示すフィールド（以下 ADDRESS という）であり（以下では、これらのフィールドをまとめて呼ぶときアドレスフィールドという）、他のフィールドは補助的に用いる。Scopus でも類似の方法による。

2-2　同定の流れ

機関同定を行うには、機関の名称や所在地を含むデータ（レコード）をひとつずつ読み込み、そのレコードと機関名辞書の名称データとのマッチングを行う。

2-2-1　前処理

表記の揺れをできるだけ吸収して同定漏れを防ぐため、入力された機関データと、照合する機関名辞書データの単語列に対し、以下の処理を行う。

(1)　文字の正規化（全角文字を半角に変換など）
(2)　ハイフン'-'で区切られた語、またはキャメルケースで表記された語の変換
(3)　主な前置詞、冠詞、接続詞、各種記号の除去（カンマ','はそのまま）。
(4)　略記辞書、ローマ字揺らぎ対応地名辞書、米語・英語対応辞書を用いて語を正規化
(5)　一部語尾の処理

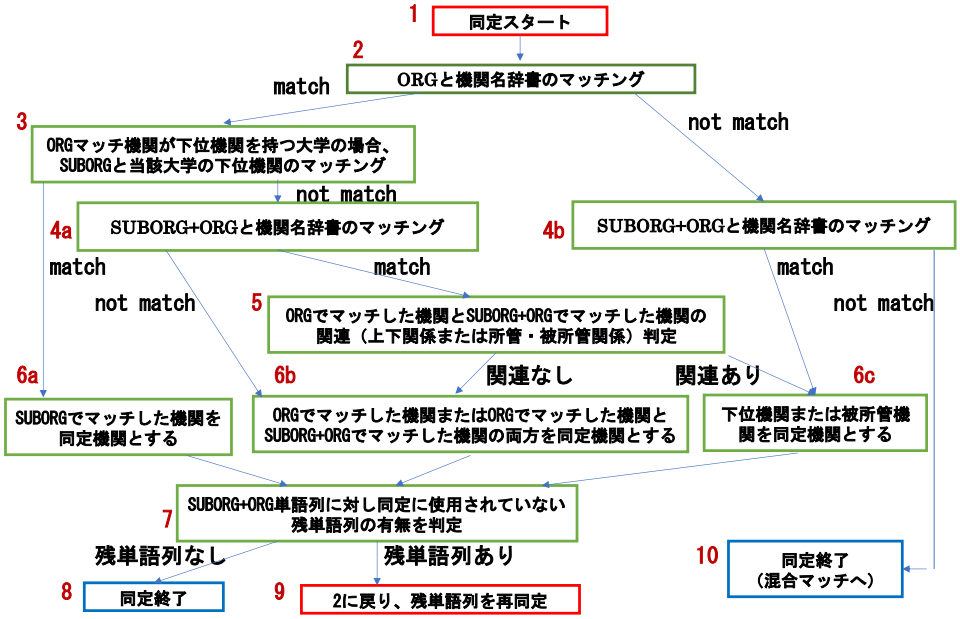
2-2-2　最長マッチ

2-2-1 において正規化を行った同定対象単語列に対し、最長の連続単語列でマッチした機関名辞書内の名称データを持つ機関を同定候補とする。概要図表 2 にその流れ図を示す。

この図のフレーム 2→3→ 4a 及びフレーム 2→4b の部分は、ORG が①下部組織を持つ大学とマッチしたか、②その他の機関とマッチしたか、③マッチしなかったか、によりその後の処置が異なることを示す。

フレーム 10 は、マッチした文字列以外に残単語列があれば、それに対しもう一度同定を行うことを示す（再帰同定）。

概要図表２　最長マッチの流れ



2-2-3　混合マッチ

WoSCC や Scopus の機関同定では、最長マッチにより 90%以上のデータが同定されるが、そこで同定できなかったデータに対して混合マッチを行う。すなわち、郵便番号マッチと曖昧マッチ（レーベンシュタイン距離が 1 以下の辞書登録機関名を探索）の 2 通りのマッチング処理を行い、その両方で同じ機関がマッチした場合に同定機関とする。

2-2-4　複数同定の場合の絞り込み

最長マッチあるいは混合マッチが終わった段階で複数機関が残った場合（同時同定）、最も確からしい機関への絞り込みを行う。主な絞り込みは、同時同定機関に上下関係、継承関係、所管・被所管関係がある場合、特定の位置付けにある機関を優先する処理である。

これらの処理を行ってもなお複数の機関が同定候補として残るときは、いずれも同定機関とする。例えば、”Div Mammalian Dev, Natl Inst Genetics, Dept Genetics, SOKENDAI”の場合、情報・システ

ム研究機構国立遺伝学研究所と総合研究大学院大学に同定される。

2-2-5　機関同定できなかったデータ

以上の処理を経ても機関同定されないデータがある場合は以下の処理を行う。

(1) セクター判定：同定対象単語列中に例えば"Co., Ltd."があれば会社セクター、"Prefectural"があれば地方自治体機関セクターに属する機関であると判定し、そのセクター名を付与する。

(2) 病院判定：同定対象単語列中に"Hospital", "Medical Center"等があれば病院であると判定し、病院フラグを"TRUE"とする。

(3) 残ったデータ：以上の判定もされなければ同定不能とする。

2-2-6　ベクトルマッチ

2-2-5 のデータに対し、オプションでこの処理を行って再同定を図る。機関名辞書の個々の収録機関を 1 つの文書と見なした TF-IDF 型単語ベクトルのファイル（ワードベクトルファイル）に、同定対象データから作った TF-IDF 型単語ベクトルをマッチさせて、最高の類似度を持つ機関が定められた閾値類似度を超す場合に同定機関とする。

機械的な同定処理は以上で終了する。

3　名寄せの要注意点とそれへの対処

この章では、名寄せを行うとき特に問題となる点と、それに対する機関名辞書、名寄せプログラムでの対応策について述べる。

3-1　下部組織の同定

3-1-1　大学の下部組織

(1) 大学の下部組織に特有の最長マッチ方式

WoSCC や Scopus での大学組織のデータは、多くの場合 ORG に大学名、SUBORG に組織名が記入されている。また、大学の下部組織の特徴として、同じ名称の学部や大学院研究科が多くの大学に存在する。従って、ORG で大学にマッチすれば SUBORG でその大学の下部組織とマッチさせることが効率的な同定に結びつく。そこで、概要図表 2 のとおり、ORG マッチで下部組織を持つ大学に同定されれば、まず SUBORG マッチにより同定された大学の下部組織とのマッチングを行い、マッチしない場合に SUBORG+ORG 同定（SUBORG と ORG を接続した単語列のマッチ）を行う。

(2) 32 大学の下部組織同定に用いるサポート辞書

1-3-2(1)に述べた 32 大学については、機関名辞書では第 2 階層下部組織を網羅的に収録している。しかし、WoSCC や Scopus のデータでは、例えば"Department of Physics, the University of Tokyo"のように、第 2 階層の"Faculty of Science"を省略した表記が屡々見られる。これに対処するた

め、32 大学の下部組織同定の際、次のサポート辞書を用いる。

- 下位機関統計辞書:機関名辞書に収録されていないが SUBORG によく出現する第 3 階層以下の組織名を、その上位の第 2 階層組織(機関名辞書に収録)に統計的に結びつけた辞書。
- ユーザー定義統計辞書:ある第 2 階層組織に関係づけて間違いないと考えられる単語列(概ねその下位の組織を表す)を収録した辞書。

ORG における最長マッチで 32 大学のいずれかが同定された場合、まず SUBORG、次に ADDRESS を機関名辞書の下部組織名とマッチさせ、マッチしなければ上記 2 つのサポート辞書と最長マッチを行う。この方法により、第 2 階層省略データのうち少なくとも過半数が同定できている。

3-1-2　大学以外の機関の下部組織

WoSCC や Scopus のデータでは、以下の理由により下部組織名抽出が難しい場合がある。

(1) 下部組織の英語名は、(a) 代表機関名の後に下位機関名を続ける、(b) 下位機関名の後に代表機関名を続ける、(c) 代表機関名を省略し下位機関名だけで表記する、のように多様で、ORG と SUBORG に分離していることが多い。

(2) SUBORG フィールドには、下部組織の名称だけでなく、下部組織の更に下の組織名や、所在地、郵便番号等のアドレス情報が付随している場合が多く、最長マッチが困難である。

これに対する主な対策は、概要図表 2 に示すように、ORG マッチの後 SUBORG+ ORG マッチを行う(2021 年度から ORG+SUBORG マッチも導入した)。また、補完策として、機関名辞書に Alias または Variant を補う方法がある。しかしながら、上記の(2)に対してはこれらの対策でも十分ではなく、現在検討を続けている。

3-2　機関の変遷

ある機関が別の機関に変遷すると日本語名は変わるが、英語名は変わる場合と変わらない場合がある。一方、機関の変遷がなくても(日本語名は変わらない)英語名が変更されることもある。また、英語名が変更されても論文等には旧名を表示する場合もある。これらに対して適切な同定を行うため、以下の対策を執っている。

(1) 主要な対策として、過去に存在した主要な機関をできるだけ機関名辞書に登録し、その変遷情報を記録する(1-2-5 参照)。

(2) 変遷前後の機関に対応する下部組織が存在する場合は、変遷前、変遷後の両下部組織を登録する。

(3) 英語名変更後の論文で旧名がよく用いられるときには、これらの旧名を Alias または Variant とする。

(4) 変遷前後で英語名が変わらない場合((3)の場合を含む)、同定対象のデータベースにある論文の発表年と、機関名辞書に記録された移行年を比較し、発表年が移行年以前であれば旧機関に、そうでなければ新機関に同定する。

3-3　英語名が同一または類似のため間違いやすい同定

日本語名が違ってもローマ字にすると同一になるため、英語名が同一になる機関がある。また、機関名が機関表記によく使われる単語のみから成る場合、複数のよく似た名の機関が存在することが多い。更に、機関の略称が機関名中によく現れる単語と一致する場合がある。このようなときに誤同定を引き起こす可能性が高いので、以下の対策を執っている。

3-3-1　機関名辞書での対応

(1) 類似の名称を持つ機関を機関名辞書に登録し、それらに使われるいろいろな英語名を収録する。例えば、国立研究開発法人理化学研究所の脳科学総合研究センター、埼玉大学の脳科学融合研究センター、玉川大学の脳科学研究所の下部組織英語名はいずれも"Brain Science Institute"である。埼玉大学、玉川大学の下部組織は機関名辞書の収録基準外であるが、理研脳科学総合研究センターへの誤同定を避けるために収録している。

(2) 問題となる英語名が Web や Scopus にほとんど出現しなければその英語名を NotUse にする。

3-3-2　混同しやすい大学のペアに対する特別措置

静岡大学と静岡県立大学、滋賀大学と滋賀医科大学のように、類似の英語名称を持つため同定が混同しやすい 15 の大学ペア(3 つ組の場合もある)に対して「特別措置機関統計辞書」を用意した。この辞書には、それぞれの大学独自の下部組織名、所在地、郵便番号等を示す単語列を収めている。ORG に対する最長マッチでこれらの大学のいずれかがマッチしたときは、この辞書を参照してより適切な大学の方に同定する。

3-3-3　特別ルールの設定

3-3-1、3-3-2 では対応が難しいケースには、個別に特別ルールを設定する。これは、一方の機関のアドレスフィールド単語列中に含まれる可能性の高い語(機関が所在する地名や郵便番号の場合が多い)を利用して、「ある単語が存在する(しない)場合はある機関に同定する(しない)」といったルールを設ける処理である。現在 32 の特別ルールを設けているが、2 つの例を挙げる。

- 清泉女子大学と聖泉大学の英語名はともに Seisen University であるが、前者は東京都品川区に、後者は滋賀県彦根市にあることを利用して識別している。
- 機関名辞書には国際基督教大学の略称として"ICU"が登録されているが、病院の集中治療室(Intensive Care Unit)を表す"ICU"がアドレスデータに含まれると国際基督教大学に同定されてしまう。このため、アドレス単語列中に"Hosp", "Med"または"Hlth"が含まれれば国際基督教大学を同定対象から外すルールを設けている。

3-4　表記の揺れ

これには、(a)正式の名称とは単語や語順が異なる表記、(b)単語の略記及び冠詞、前置詞、接続詞の省略、(c)機関・組織の略称、(d)スペル方式の違い、等があるが、主に次の2つの方法のいずれかで対処

している。

3-4-1　前処理での対応

上記の(b)に対しては、主に 2-2-1 で述べた前処理において対処する。WoSCC の所属機関データでは略記表に定められた略記が用いられるが、前処理で用いる略記辞書では、それらの他に WoSCC や Scopus に現れる不統一な略記にもかなり対応している。上記の(d)には、ローマ字揺らぎ対応地名辞書及び米語・英語対応辞書を用いる。

3-4-2　機関名辞書での対応

上記の(a)と(c)に対してはこの方法による。よく知られている略称は機関名辞書の Alias とする。毎年行う WoSCC や Scopus の名寄せ結果から、表記揺らぎのために同定洩れになったり正しく同定されなかったりしたデータをチェックし、必要と考えられる Variant を追加している。また、(b)と(d)に対しても、前処理では対処できない場合、Variant の補充を行うことがある。

4　名寄せ結果の調査・検討

WoSCC あるいは Scopus の著者所属機関データの名寄せを行った結果、機関の同定がされなかったデータが現れる。また、3 に述べたように誤同定を避けるためのいろいろな工夫をしているが、それでも少数ながら誤りは発生する。同定された結果あるいは同定されなかった結果をチェックし、同定漏れや誤同定をできるだけ低くするための対策を検討する作業が必須である。

4-1　機関同定できなかったデータの調査

2022 年 5 月に行った WoSCC データの名寄せ（1996～2021 年発表論文中の日本所属機関データが対象）では、処理された約 581.0 万レコードに対して機関同定されたのは約 549.4 万レコード、充足率は 94.6%であった。Scopus に対する充足率はそれよりやや低く 92.5%である。

出現頻度がある程度以上の未同定データに対してその理由を検討し、以下のいずれかの処置を行う。

(a) 辞書に未収録の重要な機関であるので、新たに登録する。

(b) 辞書に収録されている機関であるが、現在の機関名辞書ではマッチしない表記が今後も見られると予想されれば、機関名辞書への Variant の追加、ユーザー定義統計辞書または略記辞書への追加等を行う。稀にではあるが同定アルゴリズムの修正や特別ルールの設定を行うこともある。

(c) 特に処置は行わない。(a)または(b)を行うことによって誤同定等好ましくない影響が予想される場合、あまりにも基準から外れた表記である場合、既に存在しない等の理由により今後の出現があまり予想されない場合にはこの選択肢を採る。

機関名辞書では正解率の低下（つまり誤同定の増加）を招かないことを、充足率の上昇よりも重視しているので、未同定データの救済は誤同定の発生を招かない範囲で行っている。

4-2　主な誤同定の内容とそれへの対処

WoSCC や Scopus の 20 年分以上の著者所属機関データ（日本にある機関）は数百万件に昇り、機関同定の結果を逐一チェックする訳にはいかないので、同じ同定機関の中で ORG+ SUBORG+ ADDRESS サブフィールドが全く同じレコードをまとめ、まとめたレコードの数がある頻度以上のレコードをチェックの対象とする。単純に出現頻度だけで抽出すると大規模な機関のデータに偏るので、同定機関全体の出現頻度を考慮して、低い出現頻度の機関からもある程度が抽出されるようにする。これらに対して目視でチェックを行い、同定が誤っているデータはその理由を検討する。

2020 年度に行った WoSCC データの名寄せ（1998～2019 年発表論文中の日本所属機関データが対象）のチェックに基づく正解と誤同定の分布は概要図表 3 の通りである。

概要図表３　2020 年実施の WoSCC 名寄せにおける正解と誤同定の推定分布

同定の正誤とそのタイプ	占有率
O:正しい同定	96.69%
A1:代表機関の誤り（正解機関が辞書にあり）	0.01%
A2:代表機関の誤り（正解機関が辞書になし）	0.00%
A3:複数同定の一方の代表機関は不要（単独同定が正しい）	0.03%
B1:代表機関同定だがその下部組織が正解	0.70%
B2:同じ代表機関の別の下部組織が正解	0.00%
B3:複数同定の一方の下部組織は不要（単独同定が正しい）	0.57%
C:変遷前または変遷後の機関が正解	2.00%

エラー率は 3.3%であるが、比較的率の高い B1 と C は全く異なる機関を同定したわけではないので、これを除くと 0.6%となる。また、最も重大である代表機関に関する誤り(A1+A2+A3)は 0.04%なので、この名寄せの正確度は完全ではないがかなり高いと言える。なお、発見したエラー（それから推定される未チェックのエラーを含む）は修正しているので、公表している WoSCC-NISTEP 大学・公的機関名辞書対応テーブルや Scopus-NISTEP 大学・公的機関名辞書対応テーブルのエラー率はずっと低い。

見出されたエラーは、誤りのタイプと誤同定の理由を検討し、以下のいずれかの処置により解決を図った。これらの多くは 3 に述べた対処に含まれる。

- 新設組織の辞書への登録
- Alias または Variant の追加
- Variant を削除あるいは修正
- 所管-被所管の関係づけ
- ORG+SUBORG マッチの導入
- 特別ルールの設定
- その他のプログラム修正

これらの処理により同定の精確度は著しく改善されたので、2022 年度の名寄せでは、99.5%以上の正解率が達成されると予想している。しかし、機関の新設や改廃により新しいデータが不断に出現すること、

現在執っている対策でも完全とは言えないものがあることから、エラーの検出と対策検討は今後も必要である。

5　最後に

2011 年に整備を開始した機関名辞書は、2022 年 6 月時点で国内の約 16,300 の代表機関、約 4,300 の下部組織（非現存の代表機関約 5,700 と下部組織約 1,200 を含む）の基本的情報を収録しており、毎年更新した版を公開している。

辞書の整備とともに機関名寄せプログラムの開発を進め、WoSCC、Scopus の著者所属機関データをこの辞書の収録機関に同定する名寄せを行ってきたが、2021 年度からはその名寄せプログラムも公開している。

機関名辞書の収録カバレージ、名寄せプログラムの性能とも相当なレベルに達し、利用者からの評価を得ているが、十分な満足に至っているとは言えない。今後も利用者の拡大、データの充実、プログラムの精度向上に努めていく予定である。

本報告書で紹介した機関名辞書等の各種データは以下の HP で公開されています。
https://www.nistep.go.jp/research/scisip/randd-on-university

また、機関同定プログラムの利用を希望される方は、noip-registration[at]nistep.go.jp 宛て(機関同定プログラム担当)([at] を”@”に変更してください)にお申し込みください。お問い合わせについても、このアドレス宛てによろしくお願いいたします。