NISTEP NOTE（政策のための科学）No. 25

NISTEP における大学・公的機関名辞書の整備と名寄せプログラムの開発
－より精確な研究機関同定（名寄せ）を目指して－

2023 年 1 月

文部科学省　科学技術・学術政策研究所
科学技術予測・政策基盤調査研究センター
小野寺　夏生　　伊神 正貫

NISTEP NOTE（政策のための科学）は、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」に関する調査研究やデータ・情報基盤の構築等の過程で得られた結果やデータについて、速報として関係者に広く情報提供するため担当グループまたは著者が取りまとめた資料です。

NISTEP NOTE (Science of Science, Technology and Innovation Policy) is published as outputs of researches for "Science of Science, Technology and Innovation Policy," as well as results from data and information infrastructure, and it aims to circulate under the name of research group or author(s) as a preliminary report to the party concerned.

【調査研究体制】
小野寺夏生　文部科学省 科学技術・学術政策研究所 科学技術予測・政策基盤調査研究センター 客員研究官
伊神 正貫　文部科学省 科学技術・学術政策研究所 科学技術予測・政策基盤調査研究センター センター長

【Contributors】
ONODERA Natsuo　Affiliated Fellow, Center for S&T Foresight and Indicators, National Institute of Science and Technology Policy (NISTEP), MEXT
IGAMI Masatsura　Director, Center for S&T Foresight and Indicators, National Institute of Science and Technology Policy (NISTEP), MEXT

本報告書の引用を行う際には、以下を参考に出典を明記願います。
Please specify reference as the following example when citing this NISTEP NOTE.

小野寺夏生・伊神 正貫,「NISTEP における大学・公的機関名辞書の整備と名寄せプログラムの開発－より精確な研究機関同定（名寄せ）を目指して－」, *NISTEP NOTE（政策のための科学）*, No.25, 文部科学省科学技術・学術政策研究所.
DOI: https://doi.org/10.15108/nn025

ONODERA Natsuo and IGAMI Masatsura, "Construction of the Dictionary of Names of Universities and Public Organizations and development of the Program for Organization Name Disambiguation in NISTEP: Toward more precise and accurate research organization name disambiguation," *NISTEP NOTE (Science of Science Technology and Innovation Policy)*, No.25, National Institute of Science and Technology Policy, Tokyo.
DOI: https://doi.org/10.15108/nn025

NISTEPにおける大学・公的機関名辞書の整備と名寄せプログラムの開発
ーより精確な研究機関同定（名寄せ）を目指してー

文部科学省 科学技術・学術政策研究所 科学技術予測・政策基盤調査研究センター
小野寺夏生, 伊神正貫

要旨

科学技術・学術政策研究所（NISTEP）では、我が国における研究開発動向の機関レベル、組織レベルでの分析のための基盤として、大学・公的機関名辞書の整備と、データベース中の英語機関名データをこの辞書中の機関に名寄せするプログラムの開発を進めている。本報告書は、この機関名辞書と名寄せプログラムを活用したい人たちのために、それらの特徴、機能、利用の際に留意すること等を述べる。機関名辞書は、研究開発を行う国内機関約16,000とそれらの主要な下部組織4,000の基本的情報を収録する。名寄せに用いるための種々の英語揺らぎ名を収録していること、代表機関と下部組織の間に階層関係を付けていること、非現存機関とその継承機関の間に関係を付けていることなどが特徴である。名寄せプログラムにおいてはこれらの特徴を活かし、類似英語名を持つ機関の識別、下部組織の同定、変遷前後の機関の識別に工夫を凝らしている。

Construction of the Dictionary of Names of Universities and Public Organizations and development of the Program for Organization Name Disambiguation in NISTEP: Toward more precise and accurate research organization name disambiguation

ONODERA Natsuo and IGAMI Masatsura
Center for S&T Foresight and Indicators, National Institute of Science and Technology Policy (NISTEP)

ABSTRACT

National Institute of Science and Technology Policy (NISTEP) has been promoting construction of the Dictionary of Names of Universities and Public Organizations and development of the Program for Organization Name Disambiguation, which enable to identify an organization name description (in English) to an organization included in the Dictionary. NISTEP aims that the Dictionary and the Program serve as the infrastructure to analyze the current state and trend of research and development in Japan in micro- and meso-level. This report describes their characteristics and functions, and also issues that should be noted at their use. The Dictionary includes the basic information on about 16,000 principal organizations in Japan and their about 4,000 main subsidiary organizations. Its features are many variant organization names for disambiguation, hierarchical relation among organizations, and association of closed organizations with their succeeding ones. The Name Disambiguation Program devises precise and accurate identification of organizations, particularly in discrimination between organizations with similar names, identification of adequate subsidiary organizations, and discrimination between succeeded and succeeding organizations.

（裏白紙）

目次

概要

本編

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

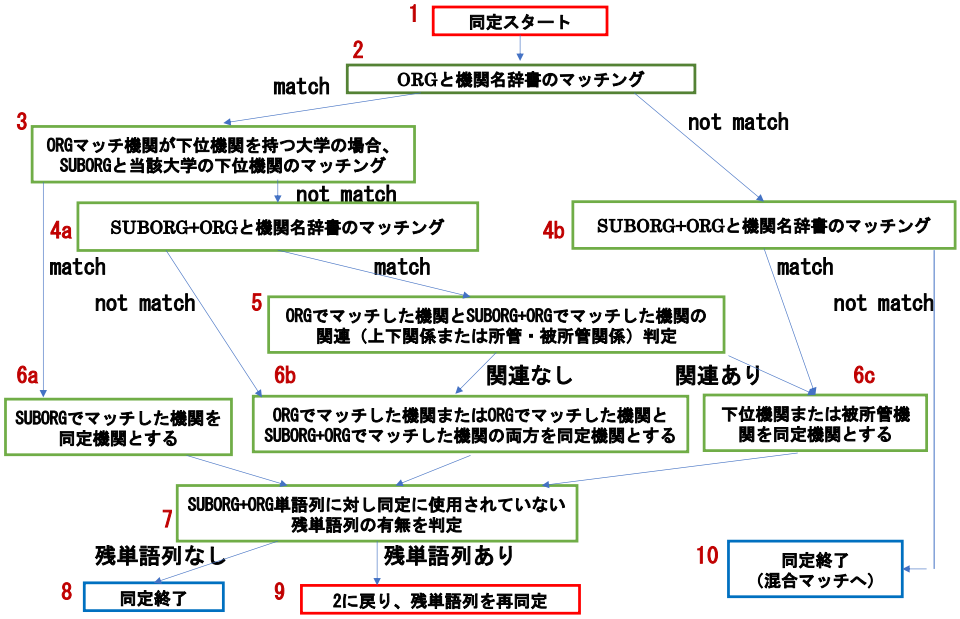
2-2-2　最長マッチ

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。

（裏白紙）

本編

（裏白紙）

はじめに

科学技術・学術政策研究所（NISTEP）では、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）」[1]の一環として、「NISTEP データ・情報基盤」の構築を 2011 年度以来進めている。その一環としての「大学・公的機関における研究開発に関するデータ整備」[2]は、個別機関（及びその組織）、セクター、国などの各レベルで研究開発の実態把握を行うための基礎となるデータの整備を行うものであり、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、この辞書を用いた機関名寄せプログラムの開発、及びそれらの活用の普及を中核的業務としている。

研究開発の動向を把握するための情報源はいろいろ存在するが、これら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おり[3]、自機関と他機関の研究比較を行うリサーチ・アドミニストレーター(RA)、政府の研究開発投資の成果や影響を定量的・構造的に分析する研究者等に活用されている。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。

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[4,5]が、2021 年度からプログラム自身の公開を開始し[6]、2022 年 1 月時点で 30 名以上の利用者がある。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。

このレポートでは、第 1 章で機関名辞書の概要、第 2 章で名寄せプログラムの仕組み、第 3 章で名寄せの際の要注意点とそれに対し辞書やプログラムでとっている対策について述べる。機関名辞書については、すでにいくつかの資料で報告している[7,8,9,10]が、この第 1 章では名寄せに関係することに焦点を当てて説明する。そして、第 2 章、第 3 章で NISTEP の名寄せプログラムでどこまでのことができるかを詳しく述べることとする。第 4 章で現在の名寄せプログラムの性能と同定洩れ及び誤同定の処理について述べ、第 5 章で未解決の課題に言及する。

この報告書で用いる独自の用語のうち複数箇所で言及するものについて、その簡単な説明、及び初出章節と詳しい説明のある章節を付録「この報告書で使用する用語について」に記した。

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。なお、大学、短大、高専、大学共同利用機関はすべての機関を網羅的に収録しているが、公的機関については、研究をほとんど行っていない機関は収録していない。

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。

第一は、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ことである。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。代表機関と下部組織の区分、及び下部組織の収録基準等については 1-3 で詳述する。

第二は、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とである。最近は、大学等、公的機関、企業を問わず、機関や組織の統廃合や改組が頻繁に行われるので、この種の情報は、機関の活動を時系列で追跡するときの困難さを軽減すると考えられる。

機関名辞書では、下部組織、非現存機関を含めて個々の機関を識別単位とし、それぞれに NISTEP 機関 ID(NID)を識別キーとして割り当てる。NID は 18 桁の固定長文字で、先頭 7 文字は "NID2012"、残りの 11 文字はランダムに発生させた番号である。

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

各機関を図表 1 に示す 17 のセクターのいずれかに分類する。このように区分を細かくすることにより、セクターレベルでの多様なデータ分析に適応できる。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。

図表 1 機関名辞書で使用するセクター

セクター番号	セクター名	備考
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	
8	国の機関	
9	国立研究開発法人等	独立行政法人、特殊法人、認可法人を含む
10	地方公共団体の機関	地方独立行政法人を含む
11	学校法人	
12	私立大学	
13	私立短期大学	
14	私立高等専門学校	
15	会社	
16	非営利団体	
17	その他の機関	日本所在の国際機関等

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、以下の点で統一を行っている。

(1) 代表機関の種別を示す接頭辞または接尾辞

- 中央省庁の施設等機関の名称には先頭に所属の省庁名を付ける。
  ［例］国土交通省国土地理院
- 国立研究開発法人等の名称には先頭にこれらの種別を付ける。公益法人等についても法人格が判る限り先頭にこれらの種別を付ける。
  ［例］国立研究開発法人産業技術総合研究所
  独立行政法人国立病院機構
  公益財団法人高輝度光科学研究センター
  ［例外］年金積立金管理運用独立行政法人
- 会社の名称には、当該機関の呼称に応じて先頭または末尾に「株式会社」、「有限会社」等を付ける。
- 大学等に対する「国立大学法人」や「大学法人」、大学共同利用機関に対する「大学共同利用機関法人」、国立高等専門学校に対する「独立行政法人国立高等専門学校機構」の種別は省く。
  ［例］東北大学
  自然科学研究機構
  久留米工業高等専門学校

(2) 下部組織名称中の代表機関名

- 下部組織の名称に対しては、代表機関名を先頭に付ける[2]。

  ［例］東京大学物性研究所

  情報・システム研究機構統計数理研究所

  社会福祉法人恩賜財団済生会横浜市東部病院

- 大学の下部組織名にその大学名を含む場合も上記の原則に従うので、次の例のように見た目にはやや不自然な表記もある。

  ［例］千葉大学千葉大学・上海交通大学国際共同研究センター

  （最初の「千葉大学」が代表機関名、それ以降が下部組織名）

- 連合大学院の連合研究科の日本語名称の先頭には通常基幹大学名が置かれるので、例えば岐阜大学連合獣医学研究科の場合次のようになる。

  (a) 基幹大学の岐阜大学での下部組織名：岐阜大学岐阜大学連合獣医学研究科

  (b) 参加大学の東京農工大学での下部組織名：東京農工大学岐阜大学連合獣医学研究科

- 大学院の研究科等では「大学院」を省略し研究科名のみを示す。但し、研究科名のない大学院や、大学院における教育・研究の改善について検討する全学的組織では「大学院」を省略しない。

  ［例］大阪大学理学研究科

  北海道大学法科大学院

  熊本大学大学院先導機構

(3) NID と日本語正式名の関係

従来、NID と日本語正式名は 1 対 1 に対応する仕様としていた。つまり、日本語正式名は識別キーの役割も持っていた。しかし、2020 年頃から、過去に存在した機関と同一の日本語正式名を持つ機関が新規に発生する現象が起こった。代表例は東京都立大学である。（旧）東京都立大学は 2005 年に首都大学東京に名称を変更したが、2020 年に再び東京都立大学になった。他にも、シチズン時計株式会社と株式会社アマダは東京都立大学と同様、一旦別の社名になった後元の名に戻った。また、ダウ・ケミカル日本株式会社は名称変更したが、その後同じ系列に同名の会社が設立されている。

そこでこの仕様を変更し、NID と日本語正式名は 1 対 1 に対応しないこととした[3]。同一の日本語正式名を持つ機関には、異なる NID が割り当てられる。たとえば、（旧）東京都立大学は NID201200599950793、（現）東京都立大学は NID201200313801084 である。同一日本語名の機関のうち現存しているのは 1 つだけなので、両者は変遷情報により区別できる。

---

[2] 2021 年度までは、大学の下部組織以外には代表機関名を省く場合もあったが、2022 年度からこのように改めた。

[3] 機関名辞書への登録システムでは同一日本語正式名の間にある区別をしているが、登録後はそれを取り除き、同定用、公開用の機関名辞書では同一名となる。

1-2-3　機関の英語名称

機関名辞書の主要な利用目的に、論文データベースその他の情報源（現在では英語のものに限る）に出現する機関名の名寄せがある。精確な名寄せは英語名称の充実にかかっているとも言える。機関名辞書では英語名称を次の 4 種に区分し、その充実に努めている。なお、学校法人（セクター番号 11）に属する機関には英語名データを付けていない。

(1) 正式名(Formal)

当該機関が定める英語名である。Web サイト等から確認できた場合のみ正式名とする。公開している最新版(version 2022.1)では、英語名データを付けない学校法人を除く 19,827 機関中正式名が存在するのは 11,064 機関である。

(2) 別名(Alias)

これには次のような名称を含む。

- 正式名と思われるが未確認の名称。会社名に多い。
- 正式名ではないがよく使われる通称。機関の旧名を含む。
- 機関の略称。奈良先端科学技術大学院大学の NAIST、国立研究開発法人宇宙航空研究開発機構の JAXA 等。
- 正式名称中の一部の語の置き換えまたは省略。国立研究開発法人国立環境研究所の正式名 National Institute for Environmental Studies に対する National Institute for Environmental Sciences、独立行政法人中小企業基盤整備機構の正式名 Organization for Small & Medium Enterprises and Regional Innovation, JAPAN に対する Organization for Small & Medium Enterprises and Regional Innovation 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。WoSCC や Scopus の名寄せの経験から補強した表記も多い。多様な表記があるので分類は難しいが(1)、(2)の変形の他に以下のようなものがある。

- 短縮語による表記。WoSCC や Scopus に使われるものが多い。国立研究開発法人情報通信研究機構に対する NATL INST INFORM & COMMUN TECHNOL など。
- 上位機関名と組み合わせた下部組織名の省略。独立行政法人日本貿易振興機構アジア経済研究所の正式名である Institute of Developing Economies, Japan External Trade Organization の上位機関部分を省略した Institute of Developing Economies など。
- ローマ字書法の揺れ、単語間のスペースの有無の違い等。九州大学の Kyusyu University（正式名は Kyushu University）、東北大学サイクロトロン・ラジオアイソトープセンターの Cyclotron and Radio Isotope Center, Tohoku University（正式名は Cyclotron and Radioisotope Center, Tohoku University）など。

(4) 非使用名(NotUse)

機関名辞書に収録するが、名寄せには使用しない名称。次の 2 つのタイプがある。

- 正式名が、同一代表機関の他の下部組織の名称と同一：東海大学では同系統の大学院研究科と学部が同じ英語正式名称を持つので、研究科では Formal、学部では NotUse とする。例えば、同大学の医学研究科と医学部の英語名はどちらも School of Medicine, Tokai University なので、前者に対しては Formal、後者に対しては NotUse とする。これにより、名寄せにおいてこの英語名とマッチすれば研究科に同定される。
- 機関の略称：多くの機関が略称を持つが、これらは短文字列で名寄せのミスを招くことがあるので、よく使われるものを除き多くを NotUse にしている。特に、一般財団法人みなと総合研究財団の WAVE や一般財団法人建設経済研究所の RICE は、他の機関の名称中の単語として含まれるので、それらの機関のデータを誤同定してしまう。

機関あたり個数は、Formal では 1 個または 0 個であるが、Alias, Variant, NotUse は 0 個以上で不定である。

図表 2 英語名称付与の例

(a) 代表機関の例：東京薬科大学

名称種別	名称
Formal	Tokyo University of Pharmacy and Life Science
Alias	Tokyo College of Pharmacy
	University of Tokyo Pharmacy and Life Science
Variant	Tokyo Univ Pharmacol
	Tokyo Univ Pharmacol & Life Sci
	Tokyo Univ Pharmacol & Pharm
	Tokyo Univ. of Pharm. and Life S.
	Tokyo University of Pharmacy

(b) 下部組織の例：東北大学多元物質科学研究所

名称種別	名称
Formal	Institute of Multidisciplinary Research for Advanced Materials, Tohoku University
Alias	IMRAM, Tohoku University
	Institute for Advanced Materials Processing, Tohoku University
	Institute for Chemical Reaction Science, Tohoku University
	Research Institute for Scientific Measurements, Tohoku University
Variant	Inst Multidisciplinary Adv Mat, Tohoku University
	Inst Multidisciplinary Res Adv, Tohoku University
	Inst Multidisciplinary, Tohoku University
	Institute of MRAM, Tohoku University
	Multidisciplinary Res Adv Mat, Tohoku University

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（例：Graduate School of Medicine, the University of Tokyo）。これは、大学下部組織の同定を確実にするためである（2-5-2 参照）。但し、下部組織名に大学名を包含する場合（XXX University Hospital な

ど）は、Alias や Variant では末尾の大学名を省略することがある。

代表機関 1 例と下部組織 1 例について、付与されている英語名称を図表 2 に示す。

1-2-4　機関の階層関係を表す情報

ある機関に対して、その直上の機関に「上位機関」のフラグが、その直下の機関に「下位機関」のフラグが与えられる。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。機関名辞書の階層関係はツリー状としている（上位の機関は複数の下位機関を持ち得るが、下位の機関は必ず唯一の上位の機関を持つ）ので、階層番号が乱れることはない。

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。また、下部組織にはその代表機関の NID を「代表機関」項目で与える（代表機関には自分自身の NID を与える）。

階層関係の情報は、名寄せ時に重要な役割を果たす。下位の機関の名称に上位機関の名称が含まれることはよくあるが、そうすると下位機関と上位機関が同時同定される。このとき、名寄せプログラムでは下位機関を優先して同定機関とする（2-5-1、2-5-4(4)を参照）。

1-2-5　機関の変遷情報

機関名辞書では、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は辞書に保持し、下記の情報項目を付記する。

(1) 移行区分

以下のいずれかにより機関が移行したとき、その機関は非現存になったとする。

- 統合：複数の機関が統合して新たな機関になった場合（別の現存機関に統合された場合を含む）
- 廃止：機関が廃止された場合
- 変更：名称が変更された場合

但し、「統合」と「廃止」の区別は困難な場合があり、多少曖昧さを含む。「変更」は日本語正式名が変更された時点で発生する。2009 年以降、法改正により財団（社団）法人は公益財団（社団）法人、一般財団（社団）法人のいずれかに移行したが、これも名称の変更なので、従来の財団（社団）法人はすべて非現存機関となった。

(2) 移行年月日（または推定移行年月日）

移行が発生した年月日を YYYY-MM-DD の形式で記載する。日が不詳で年月だけ判っている場合は DD を"00"、月日が不詳で年だけ判っている場合は MM と DD を"00"とする。移行年が確定できないが推定できる場合、推定移行年月日の項目に推定年を記載する。

(3) 継承機関

当該機関の後継機関の NID を記載する。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。なお、継承機関は 1 機関に限っているので、ある機関が 2 つ以上の機関に分離した場合は、継承機関は存在しない。被継承機関と継承機関の関係を継承関係という。

移行年月日（または推定移行年月日）と継承関係の情報は、名寄せの際に重要である。継承機関とその前身機関は、日本語名は異なるものの英語名は共通の場合が多いので、名称マッチだけではどちらが正しい同定機関であるか確定できない。その場合、辞書に示された移行年と同定対象データの存在年（該当の論文の発表年等）を比較することにより、同定機関を確定する。（2-5-4(4)を参照）

1-2-6　機関の郵便番号

機関の所在地の郵便番号を収録する。所在地が複数にわたる場合は、そのうちの少なくとも主要な郵便番号をカバーする。この情報は、名寄せにおいて最長マッチで同定ができなかった場合（2-6-1 参照）、あるいは類似の名称を持つ複数の機関がマッチしたときそのいずれかを判定する場合（2-5-3 参照）に利用される。

1-2-7　その他の情報

(1) 大学の下部組織種別

大学の下部組織には、「下部組織種別」として下記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

「学部・大学院統合」には、東京工業大学の各学院、早稲田大学の各学術院が該当する。「研究所」には、国立大学の場合は附置研究所だけを含め、公立大学、私立大学の場合は名称が「○○研究所」あるいは「○○研究センター」であるものとする。また、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）の拠点に指定された組織には、上記下部組織種別（実際には「研究所」か「全学組織」のいずれか）の後に括弧書きで「拠点」と付記する。

(2) 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）の英語名称に省庁名と研究所名を両方含んでいると、省庁（あるいはその外局）と研究所の両方に同定される。例えば National Food Research Institute, Ministry of Agriculture, Forestry and Fisheries という表記は農林水産省食品総合研究所（現存しない）を指すが、この機関のみならず農林水産省にも同定されてしまう。県の公設研究機関が名称中に "XXX Prefectural Government"を含んでいると、その県庁にも同定されてしまう。このデータ項目は、この問題を解決するために導入された。すなわち、所管する機関と所管される機関の間に所管-被所管の関係づけを行い、名寄せにおいてこの関係が付けられた対の両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（2-5-1、3-5-3 を参照）。

なお、被所管機関を外局が所管している場合、辞書では、所管機関を本省と外局の両者とすることがある。例えば独立行政法人水産総合研究センターは、農林水産省とその外局の水産庁の 2 つの所管機

関に結びついている。また、所管-被所管関係にあるすべての機関対に機関名辞書中で関係づけを行うのではなく、同定上必要な場合に限っており、現在約 270 組の関係づけがある。

(3) 外部の機関識別情報

機関名辞書の機関 ID(NID)は、次の 3 つの外部情報源の機関 ID と対応付けを行っている。

- NISTEP 企業名辞書の企業 ID:この辞書は、機関名辞書と同様、NISTEP のデータ・情報基盤事業により整備・公開されているもので、国内企業に関する沿革、所在地、規模、業種など多岐にわたる情報を含む 10,11,12)。現在、機関名辞書中の会社セクターに属する 5,087 の代表機関中 5,047 に企業 ID が付けられている。企業名辞書では各企業の詳細な情報を知ることができる。
- JISX0408(大学・高等専門学校コード):機関名辞書中の大学、短大、高専 1,678 代表機関のうち 1,510 にこのコードが付けられている。
- 科研費機関番号:各研究機関に定められた科学研究費助成事業に係る機関番号である。現在、機関名辞書中の 1,174 機関に付与されている。この番号が、機関名辞書収録機関の附属組織に対するものである場合は、その旨を注記している。

1-3　代表機関と下部組織

1-1 で機関名辞書の収録対象機関について簡単に述べたが、ここでは代表機関と下部組織の関係についてより詳しく説明する。

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:これらは代表機関とし、属するセクターは「短大」または「高専」(いずれも国立、公立、私立のいずれか)とする。
- 国立高等専門学校: 国立の高等専門学校は独立行政法人国立高等専門学校機構の下の組織であるが、個々の国立高等専門学校を代表機関とし、属するセクターは「国立高専」とする。(独立行政法人国立高等専門学校機構のセクターは「国立研究開発法人等」)
- 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(気象庁気象大学校等)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

(1) 大学の下部組織

大学下部組織のうち下記に属する組織は、階層の如何に関わらず必ず収録する。

- 附属病院
- 国立大学の附置研究所
- 共同利用・共同研究拠点及び世界トップレベル研究拠点形成プログラム(WPI)拠点に指定された組織。1-2-7(1)に述べたように、これらの組織の下部組織種別には括弧書きで「拠点」と付記する。

更に、下記の 32 大学[4]については特に下部組織収録に力を入れ、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[5]。第 3 階層であっても、多数の論文を発表している組織、多くの組織を擁する全学組織（「○○機構」等）の下部組織も収録することがある。

【下部組織を包括的に収録する 32 大学】

［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）は必ず収録する。

- 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）の主要な研究所、病院、及びプロジェクト事業
- 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

この他にも、研究開発を行っている第 2 階層、第 3 階層組織を選択的に収録する。

(3) その他の機関の下部組織

地方公共団体の機関、会社、非営利法人では下部組織の収録は比較的少なく、重要と考えられる組織を選抜している。その中でも病院は重視しており、地方独立行政法人である病院機構に所属する病院、大企業の附属病院（企業従業員以外にも公開しているところ）、非営利法人団体の病院の多くを収録している。

なお、他の機関と混同しやすい名称を持つ機関は、必ずしも機関名辞書の収録対象でなくとも、名寄せの必要上収録することがある。これについては 2-5-3(3)を参照されたい。

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

機関の種別を問わず、その新設、廃止、改組、名称の変更は始終行われる。また、英語名称や郵便番

[4] この数を拡充するための方法を現在検討中である。

[5] 第 2 階層組織が非常に広域的な組織の場合、その下に含まれる第 3 階層組織も収録対象とする。具体的には、金沢大学の各学域の下の学類及び各研究域の下の学系等、信州大学の各学域の下の学系、新潟大学の教育研究院の下の学系、早稲田大学の各学術院の下の学部、研究科等が該当する。

号も変更されることがある。そのため、不断の情報収集による更新が必要である。

(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 月に調査を行う。まず、文部科学省のホームページ(学校法人については日本私立学校振興・共済事業団の学校法人情報検索システム)にあるリストを、既存の辞書の代表機関と比較することにより、新設と廃止の機関を見出し、それらの機関の Web ページから必要な情報(廃止機関の移行年月日や継承機関、新設機関の英語名、郵便番号等)を知る。その後、個々の機関の Web ページに当たって、既存の辞書情報に変更がないか確認する。特に下部組織の変更が重要である。なお、1-3-2(1)に記した下部組織を包括的に収録する 32 大学については、大学の担当部局の協力をいただいて、調査した下部組織について確認と修正を依頼している。

国の機関、国立研究開発法人等については、毎年 1 月に政府(主に内閣官房)のホームページにあるリストに基づいて代表機関の新設、廃止を見出し、その後個々の機関の Web ページで詳細を調査する。調査の重点は大学等と同様である。

会社については、年 1 回最新版の NISTEP 企業名辞書(1-2-7(3)参照)と企業 ID によるマッチングを行い、特に変遷情報を取得する。企業 ID がマッチしない場合、企業名辞書側に調査を依頼し、企業名辞書に未登録の場合には、新たに登録を依頼して企業 ID を共有する。

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。チェック作業の詳細は 4 に譲るが、これにより、未収録の機関の追加、あるいは既収録機関への Variant の追加がなされる。

(3) その他の機会における情報の修正・更新

以上の他、機関名辞書や名寄せに関する種々の調査中、あるいは日常業務中にたまたま、情報の修正の必要に気づくことは珍しくない。これらは記録しておいて適時辞書の更新を行う。

1-4-2　機関名辞書の公開とファイル形式

上述のように、内部的には機関名辞書の更新は年 4～5 回に及ぶが、公開には確認、集計、加工等が必要なので年 1 回程度に留めている。2012 年 12 月に初版を公開し、現在の版は 2022 年 6 月に公開した [3]。また、2021 年 1 月に初めての英語版を公開し、2022 年 8 月には、現在の日本語版の英語版を公開した [13]。

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)から成り、その接続キーは NID である。しかし、このままのでは一覧性に欠けるため、これらを Excel の 1 シートに加工して公開している。元のファイルにあるうち、下記の情報は公開版には含めていない。

- 英語名称のうち揺らぎ名(Variant)

- 機関の郵便番号
- 所管・被所管関係
- 外部の機関識別情報のうち JISX0408 コードと科研費機関番号

また、公開の機関名辞書には英語名称のうち正式名(Formal)のみを載せ、別途公開する「大学・公的機関名英語表記ゆれテーブル」[14]に、正式名、別名(Alias)、非使用名(NotUse)を収録している(このテーブルには、WoSCC と Scopus に見られる主な名称表記も記載している)。

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。

1-5　収録する機関数と英語名称数

図表 3 は、2022 年 6 月時点における機関名辞書の収録機関数である。総数 20,526 機関、そのうち 16,272 が代表機関、4,254 が下部組織である。また、全機関中現存するのは 13,665 機関(代表機関 10,579、下部組織 3,086)である。

最初に公開した 2012 年 12 月時点では、全機関数 12,880(うち代表機関 10,606、下部組織 2,274)、うち現存機関は 12,340 機関(代表機関 10,108、下部組織 2,232)であった。非現存機関が大幅に増えており、この 10 年間に 6,000 以上に何らかの変遷があったことが判る。

次に図表 4 は、2022 年 6 月時点における機関名辞書の英語名称数である。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、国立研究開発法人等)ではほとんどの機関に正式名称(Formal)が与えられているが、地方公共団体の機関、会社、非営利団体の正式名付与率は低い。しかし、これらのセクターでは別名(Alias)の付与率が高く、これらの中には正式名が相当あると推定されるので、その確認が必要であろう。

図表 3　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

図表 4　セクター別の英語名称数（2022 年 6 月現在）

セクター	英語名称種別				Formal
	Formal	Alias	Variant	NotUse	存在率
1 国立大学	2,277	768	739	5	98.7
2 国立短大	24	4			92.3
3 国立高専	59	111	7	1	100.0
4 公立大学	212	58	66	4	96.8
5 公立短大	63	2		1	98.4
6 公立高専	7	1	1		100.0
7 大学共同利用機関	36	23	51	4	100.0
8 国の機関	204	94	260	11	98.6
9 国立研究開発法人等	725	609	979	29	94.9
10 地方公共団体の機関	881	505	249	23	59.7
11 学校法人	3				0.4
12 私立大学	1,319	440	445	31	96.8
13 私立短大	565	29	9	6	96.6
14 私立高専	4				100.0
15 会社	1,048	3,907	559	26	20.5
16 非営利団体	3,624	4,081	418	1,589	47.7
17 その他	13	4			100.0
計	11,064	10,636	3,783	1,730	53.9

2　NISTEP で実施している名寄せの方法

NISTEP では、機関名辞書に収録した機関への名寄せを行う NISTEP 機関同定プログラム（以下では「名寄せプログラム」というが、混同の恐れがない限り単に「このプログラム」あるいは「プログラム」ともいう）を開発している。「はじめに」で述べたように、このプラグラムを用いて 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 年度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。

- 機関名辞書に登録されている機関への名寄せであるので、国内の研究機関が対象である。
- 機関名辞書には日本語と英語の機関名データが収録されているので、原則的にはこの両方の言語による機関表記に適用することができる。しかし、現在の辞書は日本語の表記揺らぎには一部しか対応していないので、プログラムも英語表記を前提として作成している。
- 従来の主な名寄せ対象である 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、これらのデータベースに見られる表記揺れの多くは他のデータ源にも見られるので、国内機関の英語表記であれば一般に適用可能である。プログラムの公開を行ったので、今後は別の主なデータ源にも留意して開発を行いたい。
- 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。このため、下部組織とその代表機関の識別、非現存機関とその継承機関の識別には特に注意を払っている。
- 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。その結果生じる機関同定できなかったデータは、人手でチェックして対策を検討する。

この章では、このプログラムの概要及び同定の手順について述べる。機械的な処理について述べ、同定結果の調査・検討など、人手を伴う作業については第 4 章に述べる。

2-1　WoSCC と Scopus の著者所属機関データのフィールド構成

NISTEP では、WoSCC と Scopus の XML 形式のデータファイルを、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、種々の調査・分析とともに著者所属機関の名寄せにも用いている。以降に名寄せ方法について述べる前に、これらのデータベースのフィールド構成及び表記方法を、名寄せの対象となる著者所属機関データを中心に説明する。

2-1-1　著者所属機関データのフィールド構成

WoSCC、Scopus とも、XML 形式の提供データファイルでは、著者所属機関を示すデータは、機関名、下部組織名、所在地（都市名、町名、キャンパス等）、郵便番号、所属国等を示すフィールドに分割されている。両データベースとも、オンラインの表示やエクスポートファイルでは、これらのデータはひとつの

フィールドにまとめられているので、この点大きく異なる。NISTEP では、XML 形式から tsv ファイルに変換して名寄せに用いている。

(1) WoSCC ファイル

著者所属機関を示すフィールド"address"内のサブフィールドを図表 5 に示す[6]。

図表 5　WoSCC における著者所属機関データフィールド

元ファイルのサブフィールド	NISTEP での呼称	説明
organization	rs_organization	代表機関の名称
suborganization	rs_suborganization	下部組織の名称(郵便番号や所在地の一部が付記されることあり)
city	rs_city	所在する都市町村名
state	rs_state	日本の場合は都道府県名
zip	rs_zip	郵便番号
country	rs_country	所属国名(日本の場合は"Japan")
full_address	rs_address	以上の全データが入っている

このうち名寄せで主に用いるのは rs_organization(以下 ORG という)と rs_suborganization(以下 SUBORG という)、それに rs_address(以下 ADDRESS という)であり、他は補助的に用いる。なお、ORG と SUBORG のデータは、上記原則の逆になっている(SUBORG に代表機関名称、ORG に下部組織名称が入っている)場合がときどきあるので、名寄せプログラムではその点を考慮している(3-6-1 参照)。

(2) Scopus ファイル

著者所属機関を示すフィールド"affiliation"内のサブフィールドを図表 6 に示す。

図表 6　Scopus における著者所属機関データフィールド

元ファイルのサブフィールド	NISTEP での呼称	説明
organization	organization1	代表機関及び下部組織の名称(本文の説明を見よ)
	organization2	
	organization3	
city	city	所在する都市町村名
state	state	日本の場合は都道府県名
city-group	city_group	所在する都市町村名、都道府県名、郵便番号
address-part	text	論文著者が属するグループ名、プロジェクト名等
country	country	所属国名(日本の場合は"jpn")
ce_text	source_text	その他の所属機関データ。近年では、以上のほぼ全データが入っている(WoSCC の rs_address と同様)場合が多い。

[6] リプリント請求先著者の所属機関を示すフィールド"reprint address"に含まれるデータは"address"フィールドに含まれていないことがあるので、"reprint address"のデータも含む。

この表の organization1、organization2、organization3（以降 ORG1、ORG2、ORG3 という）について説明する[7]。これらのフィールドは名寄せに最も重要な情報を含むが、この 3 フィールドにすべてデータを含む場合、ORG1 と ORG2 にデータを含み ORG3 が空の場合、ORG1 のみにデータを含み ORG2 と ORG3 が空の場合がある（大雑把には 1:2:1 程度の割合）。各々の場合の各フィールドに含まれる情報は概ね以下の通りである。

(a) ORG1、ORG2、ORG3 すべてにデータありの場合：ORG3 に代表機関名称、ORG2 にその下の下部組織名称（学部名等）、ORG1 に更にその下の下部組織名称（学科名等）

(b) ORG1 と ORG2 にデータありの場合：ORG2 に代表機関名称、ORG1 に下部組織名称

(c) ORG1 のみにデータありの場合：代表機関名称と下部組織名称、あるいは代表機関名称のみ

但し、この振り分けはあくまで「概ね」であり、そうとは限らない事例が屡々ある（WoSCC における ORG と SUBORG の逆転より遙かに多い）ので、名寄せの際はそのことを考慮して柔軟な処理をしている。

2-1-2 レコード構成

WoSCC、Scopus とも、ファイルの 1 レコードは、原典（雑誌論文等）における 1 つの著者所属機関データに対応する。各レコードには、2-1-1 で述べた各フィールド（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）の他、記事 ID（WoSCC では UT、Scopus では Scopus ID）、記事内所属配列番号、書誌データ（掲載資料の名称、巻号ページ、発行年、タイトル等）、参考文献情報、被引用数等のフィールドがあるが、同じ記事の中の異なるレコードでは、アドレスフィールドデータと記事内所属配列番号以外は共通である。

原典の著者が複数の所属機関（または組織）に所属している場合、通常それらの所属機関を別々に記載するので、WoSCC や Scopus のファイルでも別レコードになる。しかし、1 つの所属機関情報の中に複数の機関が記入されることがある。その例を 2 つ示す。

(a) Inst Solid State Phys, Univ Tokyo, JST-CREST, Japan

(b) Div Mammalian Dev, Natl Inst Genetics, Dept Genetics, SOKENDAI

(a)には東京大学物性研究所と科学技術振興機構(JST)戦略的創造研究推進事業 CREST が、(b)には情報・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が含まれる。すなわちある機関（大学が多い）の研究者が JST のプロジェクト研究者にもなっている場合、総合研究大学院大学の大学院生が大学共同利用機関の研究所に所属して研究している場合である。

このように、1 つの所属機関レコードに複数の機関あるいは組織が存在しているので、NISTEP の名寄せ処理では再帰的に同定を行うこととしている（2-5-1 及び 3-7 を参照）。

[7] Scopus データベースには 4 つ以上の organization サブフィールドが存在することがあるが、その場合は最初の 3 つだけを取り出している。

2-1-3　各フィールドの充足度

2022 年度の NISTEP 名寄せ処理で用いた所属国が日本のレコードにおける、各アドレスフィールドの充足率(データが入っているレコードの割合)を図表 7 に示す。

このように、WoSCC では rs_state が約半分の充足率であるほかは、どのフィールドも空データは少ない。一方 Scopus では全般に充足率が低い。また、時期による変動を見ると、WoSCC では rs_zip の充足率が近年下がっている程度であるが、Scopus では変動が著しい。city_group の充足率は急減し、代わりに city と state のそれが急上昇している。city_group にデータが入っていれば city と state は必ず空であり、その逆も成立するので、次第に city_group を city と state に置き換えていると見られる。また、source_text は 2015 年まではほとんど空データであるが、それ以降急増し、2020-2021 年のデータではほぼ 100%充足されている。

図表 7　アドレスフィールドの充足率

[A] WoSCC

発表年	rs_org	rs_suborg	rs_city	rs_state	rs_address	rs_zip
1996-2005	100.0%	77.1%	100.0%	49.1%	100.0%	84.5%
2006-2015	100.0%	84.9%	100.0%	46.6%	100.0%	73.2%
2016-2021	100.0%	86.5%	100.0%	51.5%	100.0%	50.8%
Total	100.0%	83.2%	100.0%	48.8%	100.0%	69.2%

[B] Scopus

発表年	org1	org2	org3	city_group	city	state	text	source_text
1996-2005	100.0%	81.3%	29.1%	88.7%	2.6%	0.5%	0.00%	0.2%
2006-2015	100.0%	73.2%	23.3%	73.9%	14.9%	1.4%	0.01%	0.3%
2016-2021	100.0%	84.2%	26.0%	0.0%	89.7%	13.2%	0.01%	46.4%
Total	100.0%	78.8%	25.8%	56.9%	32.9%	4.5%	0.01%	13.6%

2-1-1 で述べたように、名寄せで主に用いるフィールドは WoSCC では ORG と SUBORG、Scopus では ORG1、ORG2、ORG3 であり、所在地フィールドは補助的に用いる程度であるが、それでもこれらの充足度が名寄せの確度に影響を及ぼすことがある。

2-1-4　名寄せを行う際の加工

WoSCC では rs_country、Scopus では country のフィールドを用いて、まず所属国が日本の機関のみを抽出する。次に、アドレスフィールドと記事 ID、記事内所属配列番号、出版年等以外の、名寄せに直接関係のないフィールドを削除する。

Scopus では更に次の処理を行う。

2-1-1 の(2) で述べたように、ORG1、ORG2、ORG3 の 3 フィールドにすべてデータを含む場合、ORG1 と ORG2 にデータを含む場合、ORG1 のみにデータを含む場合により、それらのフィールドに表記されるデータはそれぞれ(a)、(b)、(c)のようになる。このままでは名寄せアルゴリズムが複雑になるの

で、(a)の場合は ORG1 と ORG3 を逆転させ、(b)の場合は ORG1 と ORG2 を逆転させている。これにより、ORG1、ORG2、ORG3 の順に組織階層が下がる構成になる。

以下の名寄せ方法の説明では、WoSCC のフィールド名に合わせて、Scopus の ORG1 を ORG、ORG2 と ORG3 を合わせて SUBORG と読み替える。また、WoSCC の ADDRESS フィールドは他のすべてのアドレスフィールドのデータを含むので、Scopus の全アドレスフィールドのデータを合体したものを ADDRESS と呼ぶこととする。

2-2　使用するファイル

最も中心となるファイルは機関名辞書であるが、同定処理の過程で種々のサポートファイルやリストが参照される。そのうち、この章での記述に含まれるものを、初出するセクション番号([ ]内に示す)とともに挙げる。

(1) 語の正規化に用いるもの:略記辞書[2-4(5)]、ローマ字揺らぎ対応地名辞書[2-4(4)]、米語・英語対応辞書[2-4(4)]

(2) 大学下部組織の同定再現率向上に用いるもの:下位機関統計辞書[2-5-2(2)]、ユーザー定義統計辞書[2-5-2(2)]

(3) 類似名称の機関の識別に用いるもの:特別措置機関統計辞書[2-5-3(1)]、大学・短期大学のペア定義テーブル[2-5-2(1)]、特別ルール定義ファイル[2-5-3(2)]

(4) 複数の同定機関のフィルタリングに用いるもの:パターンマッチングテーブル[2-5-4(4)]

2-3　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。その手順は次の通りである。なお、特別の場合を除いて、同定処理では大文字と小文字の区別はしない。

(1) 前処理

同定対象レコード中及び機関名辞書中の語の正規化、冠詞・前置詞の除去等を行う。

(2) 最長マッチ

同定対象データの文字列に最長マッチする機関名辞書中の名称データを持つ機関に同定する。この時点で同定されたデータには同定フラグ"L"を与える。

(3) 混合マッチ(郵便番号マッチ＋曖昧マッチ)

最長マッチで同定できなかった場合、郵便番号がマッチし、かつ 1 文字違いを許容する N-Gram 文字列マッチ(曖昧マッチ)にも適合する機関に同定する。この時点で同定されたデータには同定フラグ"M"を与える。

(4) 機関同定できなかったレコードの処理

(2)、(3)のいずれでも機関の同定ができなかったデータに対して次の処理を行う。

(i) 機関が属するセクターを判別し、判明した場合同定フラグ"S"を与える。

(ii) 病院であるかどうかを判別し、そうである場合同定フラグ"H"を与える。

(iii) 以上のいずれも不明な場合同定フラグ"N"を与える。

(5) ベクトルマッチ

(4)で同定フラグ"S", "H", "N"のいずれかが与えられたデータに対し、もう一度この同定を試みる(この処理はオプションであり、公開している名寄せプログラムには含まれていない)。これは、同定対象データと機関名辞書の名称データをワードベクトルに変換し、両ベクトルの類似度が最高の機関を、その類似度が所定の閾値を越えている場合に限り同定機関とするものである。この時点で同定されたデータには同定フラグ"V"を与える。

なお、機関を示すべきデータフィールドが空白の場合は、エラーデータと見なして同定フラグ"E"を与える(この処理はレコードを読み込んだ直後に行う)。

以下、2-4～2-7 では、上記の(1), (2), (3), (4), (5)の処理についてより詳細に述べる。

2-4　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。

(0) 機関名辞書中の英語名称データはすべて半角文字で表されているので、入力データ中の全角文字を半角に変換する。また、類似する文字群を正規化する。例えば、全角のハイフン'―'は、文字コードとしていくつものパターンがあり得るが、それらをすべて半角のハイフン'-'に統一する。

(1) 入力データ中の語が(A)ハイフン'-'で区切られているとき(例えば"Radio-Isotope")または(B)キャメルケースで表記されているとき(例えば"RadioIsotope")は、(A), (B)とも(C) "Radio Isotope"と(D) "Radioisotope"の 2 通りに変換する。機関名辞書の表記も同様に処理するので、どちらでもマッチするようになる。

(2) 前置詞['of' , 'for' , 'on' , 'at', 'in', 'into']、冠詞['the']、接続詞['and']を除去。但し、これらの語が末尾にある場合、意味のある語であることがあるので削除しない(例えば、「NTT AT」という社名がある)。

(3) 各種記号['/', '&', '(', ')', '.', ':', ';', ""]及びアポストロフィも除去。但しカンマ','は同定上意味のある情報であるためそのままとする。

(4) ローマ字揺らぎ対応地名辞書を用いて訓令式で表記された地名をヘボン式に変換する(例えば、"Kyusyu"→"Kyushu")。また、米語・英語対応辞書を用いて英式綴りを米式綴りに変換する(例えば"Centre"→"Center")

(5) 略記辞書を用いて語の正規化を行う。例えば、"Science", "Sciences", "Scientific"は"Sci"に、"Pharmacy", "Pharmaceut", "Pharmaceutical", "Pharmaceuticals", "Pharmaceutics"は"Pharm"に統一する。また、会社名を表す"Corp", "Co., Ltd.", "Ltd.", "Limited"は交互に用いられており、異なる会社を混同する恐れもごく少ないため、"Corp"に統一する。

(6) 語尾の"ogical", "ogy", "ogies", "ogist(s)", "ical", "ics", "ion(s)", "ional", "ive(s)"は自動的に除去する。また、語尾が"bility"または"bilities"であれば"bil"に変換する。

2-5　最長マッチ

最長マッチとは、2-4 において正規化を行った同定対象単語列に対し、最も長く連続した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ものである。次の 2 つの理由のために複数回のマッチング処理を行う必要があるので、同定対象単語列中の単語に位置情報を与え、マッチングに成功した際にその位置を記録する。

- 2-1 で述べたように、WoSCC や Scopus のアドレスフィールドは複数のフィールドに分割されている。
- ひとつのアドレスフィールドデータに複数の機関が含まれていることがある(2-1-2 参照)。

2-5-1　概要

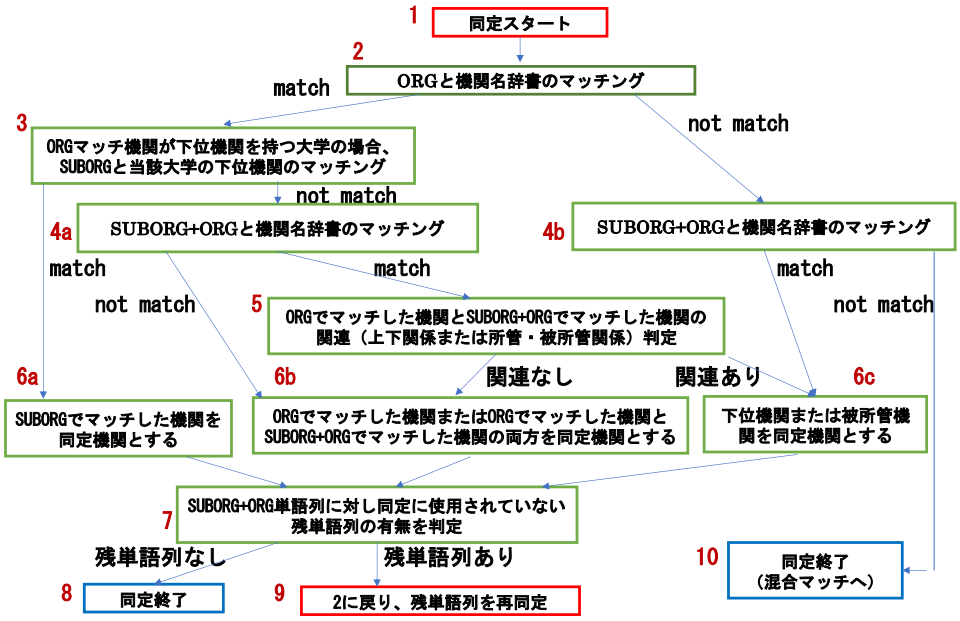
図表 8 は最長マッチ処理の流れ図である。以下、この図に沿って説明する([ ]内番号は図表 8 中の番号に対応)。

(1) ORG 単語列に対し機関名辞書を最長マッチ。マッチする単語列があればその位置を記録し[2]、マッチした機関の種類により(2)または(3)に進む。マッチする単語列がなければ(5)へ。

(2) ORG でマッチした機関のセクターが国立大学、公立大学、または私立大学で、かつ機関名辞書に下位機関を持つ場合、その下位機関と SUBORG 単語列の最長マッチを行い[3]、マッチすればその位置を記録し、マッチした下位機関を同定機関として(6)へ[6-a]。マッチしなければ(3)へ。大学の下部組織同定の方法を他の機関と区別している理由、及びその方法の詳細については 2-5-2 を参照。

(3) ORG でマッチした機関が(2)以外の場合、または(2)に進んだがそこでマッチする機関がなかった場合は、SUBORG 単語列の後に ORG 単語列を接続したもの(SUBORG+ORG と呼ぶ)に対し機関名辞書の最長マッチを行う[4-a]。この処理は、機関名が ORG と SUBORG に分離している場合に有効なことがある。マッチすればその位置を記録し(4)へ。マッチしなければ(1)でマッチした機関を同定機関として(6)へ[6-b]。

図表 8　最長マッチの流れ



(4) (1)で最長マッチした機関（ORG マッチ）と(3)で最長マッチした機関（SUBORG+ORG マッチ）の絞り込みを行う[5]。詳細は省くが、両者の間に上位・下位の関係あるいは所管・被所管関係があれば下位機関または被所管機関を同定機関として(6)へ[6-c]。それらの関係がない場合は、ORG の最長マッチ部分と SUBORG+ORG の最長マッチ部分に重なりがあれば ORG マッチした機関、重なりがなければ ORG マッチ機関と SUBORG+ORG マッチ機関の両方を同定機関として(6)へ[6-b]。

(5) SUBORG+ORG 単語列に対し機関名辞書の最長マッチを行う[4-b]。マッチすればその位置を記録し、マッチした機関を同定機関として(6)へ[6-c]。マッチしなければ(7)へ。

(6) 同定機関は決まったが、SUBORG+ORG 単語列において、同定に関係して位置を記録した部分以外の単語列が残っていないか調べる[7]。残っていなければ同定処理は終了する[8]。残っている場合、残った単語列を抽出して上記の(1)からの処理を再適用し、もう一度最長マッチを行う（これを「再帰同定」という）[9]。ここでまた同定機関が見出されれば、更に残った単語列を調べることになるが、同定機関が見出されなかった場合は(7)に行くことなく同定処理は終了する。

(7) 最長マッチでは機関を同定できなかったので、混合マッチ（曖昧マッチ＋郵便番号マッチ）に進む[10]。

再帰同定を行った結果、複数の機関が同定されることがある。また、1 サイクルの最長マッチ処理においても複数の機関が同定されることがある（同じ英語名称を持つ複数の機関がある場合等）。このような同定を「同時同定」というが、1 件のデータの最長マッチ処理が終わった時点でそれらの絞り込みを行い、最も可能性の高い機関に同定するようにしている（2-5-4 参照）。それでも 1 機関に絞りきれない場合は、残った複数の機関が同定機関となる。

なお、ここで行った SUBORG+ORG マッチに加えて ORG+SUBORG マッチを行う行程を 2021 年度に追加した。これは、ORG+SUBORG 単語列との最長マッチによって新たに同定される名称があり、それによる悪影響（誤同定）が発生しないという見通しが得られたためである。

2-5-2　大学下部組織の同定

2-5-1 の(2)で、ORG でマッチした機関が機関名辞書に下部組織を持つ大学である場合、その下部組織と SUBORG 単語列の最長マッチを行うと述べた。ここではその詳細を記す。

(1) 大学の下部組織に特有の最長マッチ方式

大学の下部組織の最長マッチ方式を、他の機関の下部組織のそれと区別している理由を述べる。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ことがある。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、精度も確度も高くかつ効率的な同定に結びつく。

なお、1-3-1 で述べたように、大学に属する短期大学部は、機関名辞書では代表機関として扱っている。しかし、これらの短期大学部の英語名の多くは属する大学名の後に"Junior College"等が付くので、WoSCC や Scopus では、大学名を表す部分が ORG サブフィールドに、"Junior College"等が SUBORG サブフィールドに分離し、短期大学部ではなく大学に同定されてしまう。これを防止して短期大学部に正しく同定されるように、辞書とは別に「大学・短期大学ペア定義テーブル」を作り、ORG で短期大学部を持つ大学にマッチし、SUBORG に"Junior College"等を含む場合は短期大学部の方に同定するようにした。

(2) 32 大学の下部組織同定に用いるサポート辞書

機関名辞書への大学下部組織の収録についての考え方は 1-3-2(1)に述べたとおりであるが、このうち 32 大学の下部組織の同定については特に注意を払っている。これらの大学については、機関名辞書では第 2 階層下部組織（学部、大学院研究科等、大学の直下の組織）を網羅的に収録し、第 3 階層以下の組織（学部の下の学科、大学院研究科の下の専攻コース等）は特別の場合しか収録していない。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、このプログラムでは 32 大学の下部組織同定の際、次の 3 つ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）に結びつけた辞書。この辞書は、同定対象ファイル（WoSCC や Scopus の大量データファイル）において第 2 階層組織名が記載されたデータから第 2 階層以外の組織名単語列を抽出して集計し、出現頻度が高く、ある第 2 階層組織に属する確率が統計的に極めて高い単語列を取り出したものである。例えば、ある大学に同定されたデータの中で、"Department of Bioscience"が結びついた第 2 階層組織がほとんど"Graduate School of Science"であればこの

辞書に取り入れられるが、"Graduate School of Science"の他に"Graduate School of Pharmaceutical Science"にもある程度結びついていれば、辞書には取り入れない。

現在の下位機関統計辞書には、次の条件のいずれかを満たす約 5,700 の単語列を収録している。

a) 出現頻度が 30 以上で、ある第 2 階層下部組織との共起率 0.8 以上

b) 出現頻度が 10 以上 30 未満で、ある第 2 階層下部組織との共起率 1

こうして作成した下位機関統計辞書を、最長マッチで 32 大学に同定されたがその下部組織同定はできなかったデータ(機関名辞書に含まれる下部組織名とはマッチしなかったデータ)に適用する(下記の(3)を参照)。

- ユーザー定義統計辞書

下位機関統計辞書は、同定対象ファイルが更新される度に作成し直すので、更新される前に存在したレコードが更新後も存在する保証はない。このため、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)を収録したものがこの辞書である。たとえば東京農工大学の場合、"Dept Anim Life Sci"は農学研究院(Institute of Agriculture)そのものを示す名称ではないが、同大学に同定された機関データにこの表記があれば、農学研究院に同定して間違いはない。同様に、"Dept Math"の表記は確実に工学研究院(Institute of Engineering)に結びつけられる。

使用目的、使用場面は下位機関統計辞書と同じである。

- 不要単語辞書

下位機関統計辞書には、時折不適切なレコードが混入することがある。SUBORG サブフィールドに記載される所在地を示す単語列が、この所在地にある主要な第 2 階層組織(A)に結びつけられることがよくあるが、この所在地にはよりマイナーな他の第 2 階層組織(B)も存在しているので、B に同定されるべきデータが誤って A に同定されてしまう。このような単語列を不要単語辞書に登録し、下位機関統計辞書が適用されないようにしている。

(3) 32 大学の下部組織同定の手順

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、次の手順でその下部組織の同定を行う。同定がなされた段階で、下記に示す「下位機関同定フラグ」を与える。

① SUBORG を機関名辞書と最長マッチ(下位機関同定フラグ S)

② ADDRESS を機関名辞書と最長マッチ(下位機関同定フラグ A)

以上でマッチした下部組織がない場合に限り以下の③と④を行う。

③ SUBORG を下位機関統計辞書及びユーザー定義統計辞書と最長マッチ(下位機関同定フラグ D)

④ ADDRESS を下位機関統計辞書及びユーザー定義統計辞書と最長マッチ(下位機関同定フラグ E)

⑤ 以上の処理で同じ下部組織が重複しておれば、より上位の下位機関同定フラグを残す。重複を除いても複数の下部組織があればそのままとする。下部組織へのマッチがない場合は、ORG でマッチ

した大学を同定機関とする。

2-5-3　間違いやすい同定を防ぐための特殊処理

(1) 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の正式名はそれぞれ"Shizuoka University"、"University of Shizuoka"である。また、滋賀大学、滋賀医科大学の正式名はそれぞれ"Shiga University"、"Shiga University of Medical Science"である。これらの名称が ORG サブフィールドに正しく記載されていれば、最長マッチにより問題なく同定されるが、静岡大学と静岡県立大学の場合、間違って表記されていることがある(データ源である雑誌論文の所属機関表記が間違っていることもある)。この 2 つの大学はどちらも静岡市駿河区にあり(静岡大学は浜松市にもキャンパスがあるが)、郵便番号の最初の 3 桁は共通で、やや似通った学部が存在するので、大学名の表記が間違っていると識別は極めて困難である。また、滋賀医科大学の名称が、ORG サブフィールドに"Shiga University"、SUBORG サブフィールドに"Medical Science"と分離していることがよくあり、そうすると滋賀大学に同定されてしまう。

これによる誤同定を防ぐために、類似名称を持つ 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。下位機関統計辞書と同様、この辞書も同定結果ファイルの統計処理によって作成する。

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、特別措置機関統計辞書でその大学とそれとペアをなす大学の項を参照し、SUBORG、ADDRESS、ZIP(郵便番号が記入されたフィールド)にその単語列を含む方の大学に同定する。例えば、ORG で静岡大学か静岡県立大学のいずれかにマッチした場合は特別措置機関統計辞書に"Fac Eng", "Res Inst Elec"等があれば静岡大学に、"Sch Pharm Sci", "Grad Sch Nutr Environ Sci"等があれば静岡県立大学に同定する。

特別措置機関統計辞書の参照によって ORG でマッチした大学を別の大学に変更して同定した場合(スワップ同定)は、同定機関交換フラグとして"P","Q",または"R"を与える。

(2) 特別ルール

異なる機関が同一の英語名称を持つ場合、措置を講じなければ両者に同時同定されてしまう。また、ありふれた単語からなる名称を持つ機関は、それらの単語を含む別の機関(機関名辞書に収録されていないものを含む)を誤って同定しやすい。特別ルールは、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場合はある機関に同定する」あるいは「ある単語が存在しない場合はある機関に同定しない」といったルールを設ける処理である。これにより同定候補となった機関の間に継承関係がある場合は、2-5-4(4)に述べる方法(機関名辞書に記されている移行年と同定対象データの発表年の比較)によりいずれかに同定する。特別ルールによってある機関に同定された場合、同定フラグ"R"を与える。

現在、32 の特別ルールを設けているが、2 つほど例を挙げて説明する。

清泉女子大学と聖泉大学の英語名はどちらも Seisen University である。最長マッチでこのフレー

ズにマッチしたときどちらか判別できないが、清泉女子大学は東京都品川区に、聖泉大学は滋賀県彦根市にあることを利用し、アドレスフィールド(郵便番号を含む)に"Tokyo", "Shinagawa", "1410022", "1418642"のいずれかがあれば前者に、"Shiga", "Hikone", "5211123"のいずれかがあれば後者に同定するという特別ルールを適用する。(これらのいずれの語も見出せないときは 2 つの大学に同時同定する。)

もう一つの例は、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(2012 年以前は財団法人)に関するものである。この英語名は Institute for Environmental Sciences であるが、この名はかなりありふれており、同じ単語列を含む地方自治体や大学等の研究所が、機関名辞書に登録されていないものを含め多数存在する。従って、"Institute for Environmental Sciences"に最長マッチしたからといってこの機関に同定することは危険である。そこで、アドレスフィールドに"Aomori", "Rokkasho", "0393212", "Radioecol", "Radiobiol"のいずれかが存在する場合のみこの機関に同定する(同定対象の論文の発表年が 2012 年以前なら財団法人、2013 年以降なら公益財団法人)という特別ルールとしている。(これらが存在しない場合は通常の最長マッチに戻る。)

32 の特別ルールを目的別に分類すると次のようになる。詳しくは、第 3 章の該当節(以下、→の後に示す)を参照されたい。

- 異なる機関の略称が一致する場合、またはある機関の略称が他機関の名称中の語と一致する場合の識別→3-1-3
- 関係を持つ機関同士、または同一機関内の継承関係にない組織同士が同一名称を持つ場合の識別→3-4-2
- 英語名の綴りが一致する無関係の機関同士の識別→3-4-3
- 類似名称を持つ異なる大学の識別→3-5-1
- ありふれた語のみからなる機関名のため他機関と区別し難い場合の識別→3-5-2
- 事業の共同運営等を行っているため名称の共通部を持つ機関間の識別→3-5-3

(3) 類似名称機関の機関名辞書への収録

似た英語名称を持つ機関を誤って同定することを避けるためには、前述の特別ルール以外に、類似した名称の機関を(機関名辞書の収録基準外であっても)収録する方法がある。例えば、自然科学研究機構分子科学研究所の英語名は Institute for Molecular Science であるが、これとごく類似した名称の付属研究所を持つ機関は、愛知医科大学の分子医科学研究所(Institute of Molecular Medical Sciences)等多数存在する。また、国立研究開発法人理化学研究所に脳科学総合研究センター(Brain Science Institute)という組織があるが、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英語名も同じ"Brain Science Institute"である。愛知医科大学、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、分子科学研究所や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。同様に、旧・国立公衆衛生院の英語名称 The Institute of Public Health は、多くの地方自治体の衛生研究所の名称にも付けられている(愛知県衛生研究所; Aichi Prefectural Institute of Public Health 等)。これらの名称は最長マッチで Institute of

Public Health より優先するので、このような名を持つ機関をできるだけ登録することにより、誤同定を防ぐことができる。

2-5-4　複数同定の場合の絞り込み

最長マッチでは、一部同定候補の絞り込み処理はあるものの、基本的になるべく多くの同定候補を抽出するようにしている。最長マッチが終わった段階で複数機関が残った場合、最も確からしい機関への絞り込みを行う。以下にそのいくつかの方法を示すが、最も重要なのは(4)のパターンマッチング処理である。

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。複数の機関が同定される理由は 2-1-2 で述べたとおりである。

(1) 同一機関の重複削除

ORG と SUBORG に同じ文字列が入っていた場合など、同一の機関が同定候補として抽出される場合がある。このときはその重複を削除する。

(2) 郵便番号マッチ

同定されたデータに郵便番号が含まれていれば、機関名辞書の郵便番号と比較し、マッチする方を同定機関とする。

(3) 連合大学院絞り込み特別ルール適用

多くの連合大学院研究科の英語名は、例えば"United Graduate School of Veterinary Science"（岐阜大学連合獣医学研究科）のように大学名は含まれない。従って、連合大学院研究科に所属する研究者が論文を発表する場合に正しい組織名を記述すれば、同定も正しくなされるであろう。しかし実際には、関係している複数の大学が様々な順番・形式で表記されることが多く、しばしばそれらの大学の同時同定になってしまう。極端な場合、大阪大学・金沢大学・浜松医科大学・千葉大学・福井大学連合小児発達学研究科は英語名も United Graduate School of Child Development, Osaka University, Kanazawa University, Hamamatsu University School of Medicine, Chiba University and University of Fukui とすべての参加大学が含まれており、表記も多種多様である。そのため、表記に"United Graduate School"が含まれる場合、次の方法により著者が所属する大学を推定し、その連合大学院研究科のみに同定する。

- データに郵便番号が含まれるときは、それに合致する大学に同定
- 郵便番号が含まれないときは、ORG マッチした大学に同定

(4) パターンマッチングによる絞り込み

同時同定の多くは、上下関係や継承関係がある機関である（下位機関名の中に上位機関名が含まれていたり、変遷しても英語名は変更しないままであったりするため）。そのため、同定候補に挙がった機関の関係を調べ、上下関係や継承関係のパターンに合致した際に特定の位置付けにある機関を残すような処理が有効である。このときの上下関係には、大学・短期大学部ペアに設定される仮想的な上下

関係（2-5-2(1)参照）を含む。

単純な場合、同時同定された 2 つの機関 A と B の間に A が B の下位機関という関係があれば A を同定機関とする。また、C が変遷してその継承機関が D である場合は、機関名辞書に記されている移行年と同定対象データの発表年を比較し、発表年が移行年以前なら C に、移行年より後なら D に同定する。発表年と移行年のいずれかが不明の場合は新しい方の D とする。

より複雑なパターンについても、概ねこの 2 つの原則（上下関係は下位機関を優先、継承関係は発表年と移行年の比較）に従って選定を行う。例えば、A の継承機関が B、A の下位機関が a、B の下位機関が b という関係があり、A と B の英語名及び a と b の英語名は同じであるが、このうち a は機関名辞書に登録されていないとする。すると、あるデータ表記に対して A, B, b が同時同定されることがあり得る。このときはまず A→B の移行年とデータの発表年を比較し、発表年が移行年以前であれば A が同定機関となる。そうでない場合、または発表年と移行年の比較ができない場合は、より新しい機関かつ下位機関優先の原則から b に同定される。

やや異なる場合として、A と B の英語名が異なるが似ているとすると、「A, a」というデータ表記は恐らく a を指しているが、a が機関名辞書にないので、A と b に同時同定されるであろう。しかし、このデータの発表年は A の存在期間にあるであろうから、A に同定されるという妥当な結果になる。但し、発表年のデータがなければ、より新しい機関である b に同定されてしまう。

このようにいろいろな場合が起こり得るので、継承関係と上下関係のパターンに応じてきめ細かく同定機関を選別できるようにし、現在、最大 4 機関が関係づけられた 31 のパターンに対するルールが設定されている。

(5) 2 機関比較法による絞り込み

(4)で述べたパターンマッチングは、直接の継承関係とそれらの上下関係までにしか対応していない。しかしながら、何段階もの変遷を経ているが英語名は同一またはほぼ同一の機関が少なくない。国立研究開発法人の研究所には特にこのような例が多く、複数の同定がされるがパターンマッチングでは絞り切れないことがある。これに対応するため、複数機関が同定候補として残ったとき、2 機関ずつのペアを順次比較し、変遷年、上下関係、所管被所管関係から絞り込む方法をとっている。

2-6　混合マッチとベクトルマッチ

WoSCC や Scopus の機関同定では、2-5 で述べた最長マッチにより 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は混合マッチにより、それでも同定されない場合はベクトルマッチにより処理される。2-3 で述べたように、ベクトルマッチは 2-7 の処理の後で行われるが、ここでは説明の順序を逆にする。

2-6-1　混合マッチ

ここでは、郵便番号マッチと曖昧マッチの 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。郵便番号マッチでは、同定対象データ中に郵便番号が含まれる場合、そ

れと機関名辞書中の郵便番号データをマッチさせる。曖昧マッチでは、マッチ対象の表記（単語列ではなく文字列）に対し、辞書登録機関名においてなるべく長い文字列長で N-Gram（N=3）スコアが高く、レーベンシュタイン距離が 1 以下のもの（1 文字の欠落や誤字）を探索する。

当初は、この 2 つのマッチングを独立して行っていたが、いずれもかなりの誤同定が発生することが判った。曖昧マッチでは、辞書未登録の簡単な名称の会社が 1 字のみ異なる辞書中の会社に同定されたり、機関接尾辞（"Co., Ltd."、"Corp"、"Inc"等）の誤表記のため別の会社に同定されたりするケースが多かった。郵便番号マッチでは、大きな大学や研究機関の敷地内にベンチャー的な機関が所在する場合や、研究団地のビル内に多くの企業が同居している場合に誤同定が起こりやすかった。そこで 2014 年度に両者を合体して混合マッチとし、両者の結果が一致するときに同定が成立するようにした。

曖昧マッチにはある程度処理時間を要するので、実際には、まず郵便番号マッチを行い、そこでマッチ機関が得られた場合にのみ曖昧マッチを行う。曖昧マッチにおいても、最長マッチの場合と同様、2-5-1 に沿った流れでマッチングを行い、再帰同定、複数機関同定の場合の絞り込み処理も行う。

2-6-2　ベクトルマッチ

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、最高の類似度を持ち、かつ定められた閾値類似度を超す機関を同定機関とする。最高類似度が閾値を越す機関がなければ同定はされない。

ワードベクトルファイルのデータの TF-IDF 計算には、機関同定結果のファイルのうち最長マッチによる同定がなされたレコードを用いる。更に、機関同定される回数が少ない機関にも対処するため、機関名辞書の英語名称および郵便番号も追加集計する。また、マッチング時にスコアの補正を行うために、同定結果ファイルにおける出現期間と機関名辞書の変遷年も追加する。

2-7　機関同定できなかったデータ

以上の最長マッチ、混合マッチ、ベクトルマッチの処理を経ても機関同定されないデータがある（WoSCC、Scopus では 5～7%程度）。これらについては以下の処理を行い、それぞれに同定フラグを付ける。

(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できる。「セクター識別辞書」によりこのような判定がされれば、そのセクター名と同定フラグ"S"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される。「病院識別辞書」によりこの判定がされれば、病院フラグを"True"とし、同定フラグ"H"を付与する。

(3) 残ったデータ

以上の判定もされなければ同定不能とし、同定フラグ"N"を付与する。

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

第 1 章において機関名辞書の構成について、第 2 章においてこの辞書に基づく機関名寄せの方法（主に WoSCC 及び Scopus の著者所属機関データを対象として）について述べてきた。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。従って、1、2 の内容と重複するところがあるが、その場合は 1、2 の該当箇所を参照して、なるべく記述の重複を避けている。

以下に述べる名寄せプログラムでの対応は、前処理及び最長マッチで行われる処理を述べているが、これらの処理で同定ができなかった場合、曖昧マッチあるいはベクトルマッチ（2-6 参照）により同定される可能性がある。

ここでも主に WoSCC と Scopus のデータを念頭に置いているが、多くの場合、他のデータベースやリストにおける日本機関の英語名データに適用できる。

3-1　表記の揺れ

機関名の表記揺れは、次の 5 つのパターンに類別化される。

① 正式の名称とは単語や語順が異なる表記
② 単語の略記及び冠詞、前置詞、接続詞の省略や書き換え
③ 機関・組織の略称
④ スペル方式の違い、ミススペル
⑤ 機関の旧名の表記

例えば、奈良先端科学技術大学院大学の英語正式名は Nara Institute of Science and Technology であるが、次のように表記されることもある。

- Nara Advanced Institute of Science and Technology（①の例）
- Nara Inst Sci Tech（②の例）
- NAIST（③の例）
- Nara Institute of Sciences and Technologies（④の例）

以下、①～④について事例を挙げるとともに、機関名辞書や名寄せプログラムで執っている対策を記す。⑤については 3-2 で述べる。

3-1-1　正式の名称とは単語や語順が異なる表記

【問題点】

ほとんどの大学の英語名は、XXX University か(The) University of XXX の 2 つに大別される。しかし、実際にはこれを誤記する例が少なくない。東京大学の正式英語名は The University of Tokyo であるが、Tokyo University という表記も 1%程度存在する。逆に京都大学は Kyoto University が正しいが、University of Kyoto も存在する。これが、語順が異なる表記の代表例と言える。

名称中の一部の単語が入れ替え、挿入、除去された表記はいろいろある。東京農工大学（正式名

Tokyo University of Agriculture and Technology)の例では、Tokyo Noko Univ、Tokyo Univ Agr Eng などがある。これらは東京農工大学であると推定できるが、Tokyo Univ Agr と誤記されると、これは東京農業大学の正式名なのでそちらに同定されてしまう。

医科大学では、XXX Medical University、XXX Medical School、XXX Medical College の表記が混在することが多い。埼玉医科大学の英語正式名は Saitama Medical University であるが、旧名の Saitama Medical School が今でもよく使われており、Saitama Medical College も少数ながら使われている。自治医科大学(正式名 Jichi Medical University)も同様である。また、東京慈恵会医科大学は Jikei University School of Medicine が、獨協医科大学は Dokkyo Medical University が正式名であるが、それぞれ、Jikei University、Dokkyo University School of Medicine も相当の頻度で使われている。

会社の接尾辞である"Corp(oration)", "Co., Ltd.", "Ltd.", "Limited"は、会社によっていずれかに決められているが、しばしば混用される。但し、これらと"Inc(orporated)"は系統が異なり、全く異なる会社の名称が"Corp"と"Inc"を除けば同じという例が見られる。

【対策】

この種の表記揺らぎへの対策は、機関名辞書への Variant の充実につきる。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。しかし、他の機関に誤同定される恐れのある表記(東京農工大学を Tokyo Univ Agr と誤記した上述の例など)を追加することはできない。

会社の接尾辞の揺れについては、2-4(5)で述べたように、"Corp", "Co., Ltd.", "Ltd.", "Limited"を同一視することにより対処している("KK"と"K K"も同様)。大企業では、これらの接尾辞を省略した Shiseido、Toray などの表記もよく見られる。これらは誤同定の恐れがほとんどないので Variant としているが、例えば Hitachi は日立市にある無関係の機関にマッチすることがある。従って、接尾辞を省略した Variant の登録には慎重を期している。

3-1-2　単語の略記、及び冠詞、前置詞、接続詞の省略や書き換え

【問題点】

(1) WoSCC の場合

WoSCC の所属機関データでは、University→Univ、Institute→Inst、Medicine (Medical) →Med、Science (Scientific) →Sci などの略記が用いられる。これらは略記表に定められているので表記揺れの問題はあまり起きないが、たまにそれ以外の語が略記されており、そのため同定漏れが起きることがある。また、省略のし過ぎと思われる略記がある。代表的なものは、Electric(ity)と Electronic(s)がいずれも Elect と略記されることであり、このため異なる機関の区別ができなくなることがある。

WoSCC では、冠詞(the)、前置詞(of, for, on など)は全て省略されるので、元の論文でこれらの間違いがあっても(of と for の取り違え等)それほど問題にはならない。ピリオド(.)やハイフン(-)も一般に省略されるが、ハイフンの両側の語を詰める場合(Electro-Commun→Electrocommun 等)と開ける場合

（Bio-Appl→Bio Appl 等）があり、注意が必要である。

(2) Scopus の場合

Scopus では、原則として所属機関データの省略をせずフルスペル表記するので、それが正式名である限り（多くはそうである）名寄せ上の問題はない。しかし、古いデータ（論文発表年が 2006 年以前、それ以降も少数だがあり）では語の略記がなされており、その略記法が統一されていないため名寄せに困難を伴う場合が多い。例えば、Agriculture または Agricultural に対し Agr, Agri, Agric と種々の略記がなされている。また、University を U.、Technology を T.といった極端な略記もある。

【対策】

これらの表記揺れの多くについては、入力データと機関名辞書名称データのマッチングに先立つ前処理（2-4 参照）において対処する。WoSCC で指定されている略記（語尾の"-ogical"等の省略を含む）、それ以外によく使われる略記は略記辞書（2-4 の(5)、(6)を参照）に採り入れている。また、2-4 の(2)と(3)で述べたように、多くの前置詞や冠詞、特殊記号は削除した上で名寄せを行う。Scopus の古いデータにある略記もできるだけ略記辞書に含めており、Agri, Agric, Agriculture, Agricultural はすべて Agr に正規化される（但し、Agrobiological は Agrobiol に正規化され、Agr と混同しないようにしている）。しかし、University に対する U.のように極端な略記には対応していない。

3-1-3　機関・組織の略称

【問題点】

大学の場合、産業医科大学(University of Occupational and Environmental Health)に対する UOEH、奈良先端科学技術大学院大学 (Nara Institute of Science and Technology)に対する NAIST などはよく知られている。しかし、それほどには知られていない略称もある（大学の下部組織の略称など）。

厄介な問題は、略称が他の機関の名称中の語に一致する場合である。例えば、国際基督教大学(International Christian University)の略称 ICU は、病院の集中治療室(Intensive Care Unit)を表す ICU と同一の綴りである。WoSCC や Scopus の機関データには、集中治療室を示す ICU が含まれることがあるが、何も対策を講じなければこれらは国際基督教大学に同定されてしまう。また、一般財団法人みなと総合研究財団の略称 WAVE や一般財団法人科学技術振興会の略称 FAST などは、Wave や Fast を名称中に含む機関に誤同定される。

一方、異なる機関が同じ略称を持つ場合もある。自然科学研究機構核融合科学研究所(National Institute for Fusion Science)、NARO の花き研究所(National Institute of Floricultural Science)、鹿屋体育大学(National Institute of Fitness and Sports in Kanoya)はいずれも NIFS の略称を持つので、名寄せ対象データに NIFS と表記されている場合、このいずれであるかを判定する必要がある。

【対策】

略称が他の機関に誤同定される可能性がないと考えられる場合は、機関名辞書に Alias として登録す

る。

他の機関に混同される恐れがあり、その略称が WoSCC や Scopus に屡々現れる場合は特別ルールを設定する(2-5-3(2)を参照)。上述の ICU や NIFS はこの方法で対処している。ICU の場合は、アドレス単語列中に"Hospital", "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し、含まれていなければ国際基督教大学に同定する。単語列に NIFS が見出される場合は、それ以外の語に含まれる地名、郵便番号、機関名中の特徴語により、上記に挙げた 3 機関のいずれかに同定する(NARO の花き研究所が同定候補になった場合には、更に論文発行年により、独立行政法人か国立研究開発法人かを定める)。

一方、その略称がそれほど用いられない場合は機関名辞書における名称種別を NotUse にする(1-2-3(4)参照)。この場合は、該当機関がその略称で表記されている場合同定漏れとなるが、誤同定の防止を優先させる原則からやむを得ない。

3-1-4　スペルの違い

【問題点】

単純なスペルミスは除いて、以下のような表記揺れがある。

(1) ローマ字書式の揺れ

九州大学の英語正式名はヘボン式の Kyushu University だが、訓令式の Kyusyu University という表記もある。工学院大学は、Kogakuin University(これが正式名)の他、Kohgakuin University、Kougakuin University とも表記される。

(2) 米語式綴りと英語式綴り

米語式の Center が多いが、英語式の Centre とする機関もある。語尾の"-or"と"-our"、"-ization"と"-isation"も同様である。

(3) 単数形と複数形の揺らぎ

東京薬科大学の正式英語名は Tokyo University of Pharmacy and Life Sciences であるが、Tokyo University of Pharmacy and Life Science も無視できないほど使われている。このように、Science と Sciences が混同されている例は他にもかなりある。奈良女子大学の Nara Women's University、東京女子医科大学の Tokyo Women's Medical University の"Women's"を"Woman's"と誤記する例も見られる。

(4) 単語間のハイフン挿入あるいはキャメルケース表記

2-4(1)で述べたように、Radio Isotope は Radio-Isotope、RadioIsotope（このような表記をキャメルケースという)、あるいは Radioisotope とも表記される。

(5) その他

特殊であるが無視できない表記ゆれに、府立大学や県立大学の府県を示す単語のゆれがある。たとえば京都府立医科大学は、正式名である Kyoto Prefectural University of Medicine

の"Prefectural"が"Prefectual"、"Prefecture"などに置き換わって表記される(この他に Pref、Prefect と略記した表記もある)。

【対策】

上記(1)～(5)に対して、多くは 2-4 で述べた前処理が有効である。

- 上記(1)への対策

ローマ字揺らぎに対応させた地名辞書を作成し、Kyushu と Kyusyu 等の対応関係を示している(2-4(4)参照)。

- 上記(2)への対策

米語・英語対応辞書を用意し、Center と Centre 等の対応関係を示している(2-4(4)参照)。また、略記辞書(2-4(5)参照)には、Behavior と Behaviour、Organization と Organisation など、数多くの米語綴りと英語綴りの対応関係を収録している。

- 上記(3)への対策

これに対しても多くは略記辞書により対応している。Science と Sciences は Sci に正規化される。Bioscience と Biosciences 等、頻出する複合語も収録されている。また、語尾の"-ogy", "-ogies"や"-ive", "-ives"の省略により統合される場合もある(2-4(6)参照)。

- 上記(4)への対策

2-4(1)に述べたとおりである。

- 上記(5)への対策

略記辞書により、Prefectural、Prefectual、Prefecture はすべて Pref に変換される。

- Variant の補充

(1)～(5)を通して、上記の方法では対処できない場合、機関名辞書に Variant を補充する方法がある。ハイフンで区切られた語やキャメルケース表記の語は 2-4(1)の方法で対処できるが、例えばある機関の正式名称が"Radio Isotope"という語を含み、これに対し入力データが"Radioisotope"と表記されている場合、逆に正式名称中の語が"Radioisotope"で入力データ中の語が"Radio Isotope"である場合には、この方法ではマッチすることができない。その可能性が高い場合、機関名辞書に Variant を追加して対応する。東北大学サイクロトロン・ラジオアイソトープセンターの正式名は Cyclotron and Radioisotope Center, Tohoku University であるが、Cyclotron and Radio Isotope Center, Tohoku University を Variant としている。

但し、曖昧性の高い Variant の追加は、他の機関を誤同定する可能性があるので、そのことを念頭に置いて慎重に行うことが必要である。

3-2　機関の変遷

【問題点】

いくつかの機関の統合、機関の吸収合併、ある機関が廃止されて別の機関に改組、単なる名称変更など、機関は常に変遷する。下部組織の変遷は代表機関より更に甚だしく、大学でも、国立研究開発法人

等の公的機関でも、毎年かなりの組織変更が行われる。WoSCC や Scopus 等の書誌データベースの収録文献の発表年は長期間にわたるので、機関名寄せでは現存しない機関の識別も必要である。また、変遷前後の機関を関係づけたいことが多いので、変遷情報（その時期や継承機関）を把握することが重要である。

機関名辞書では、日本語正式名の変更をもって機関（組織）の変遷と見なして新しい機関を登録し、旧機関との間に変遷情報を付けているが、このとき、英語名も変わる場合と英語名は変わらない場合がある。文部科学省科学技術政策研究所は 2013 年に文部科学省科学技術・学術政策研究所に名称を変更したが、英語名は National Institute of Science and Technology Policy のままである。また、短期大学から 4 年制大学への移行、国立研究機関から独立行政法人への改組の際にも、旧英語名がそのまま保存されていることがある。後者の例として、文部科学省の宇宙科学研究所は、2003 年に独立行政法人（2015 年から国立研究開発法人）宇宙航空研究開発機構に統合されてその宇宙科学研究所となったが、組織の英語名は Institute of Space and Astronautical Science で変わらない。このような場合、名寄せではどの時期の機関に同定すべきか難しい。また、英語名が変わった場合も、論文等では旧名で表記されることもあるので、旧機関に同定すると誤ることになる。

一方、日本語名称は変わらないが英語名が変更されることもある。Tokyo College of Pharmacy は、東京薬科大学(Tokyo University of Pharmacy and Life Sciences)の旧英語名である。東京理科大学、埼玉医科大学、自治医科大学の英語正式名はそれぞれ Tokyo University of Science、Saitama Medical University、Jichi Medical University であるが、旧名の Science University of Tokyo、Saitama Medical School、Jichi Medical School も現在でも使われている。

【対策】

(1) 機関名辞書への変遷情報の記載

機関の変遷への主要な対策は、過去に存在した主要な機関・組織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ことである。1-2-5 で述べたように、変遷情報は、移行の区分（統合、廃止、名称変更のいずれか）、移行年月日（月日まで判らないときは移行年）、及び移行後の継承機関（存在する場合）である。

(2) 変遷前機関への下部組織登録

機関名辞書では、変遷後の機関の下部組織を登録しているのに、変遷前の機関の対応する下部組織を登録していない場合がある。この場合、両下部組織の英語名が同一であると、本来は変遷前の組織を指すデータが変遷後の組織に同定されてしまう。従って、変遷前後に対応する下部組織が存在する場合は、できるだけ旧下部組織も登録するようにしている。

(3) 機関名辞書へ旧名の収録

上述したように、英語名称は機関の変遷（すなわち日本語正式名の変更）の際だけでなく、日本語名が変わらなくても変更されることがある。いずれの場合も、その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の正しい同定

WoSCC や Scopus には長期間にわたる論文が収録されているので、変遷前後で英語名が変わらない場合、同定された機関が何時の時点に存在した機関に相当するか判断がつかない。英語名称が変更されたとしても、新しい論文に変遷前の名称が使用されていると同じ問題が起こる。このときは、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（または推定移行年）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（移行年または発表年のデータがない場合は新しい方の機関に同定する）。詳しくは 2-5-4 の(4)及び(5)を参照されたい。但し、論文の公表は投稿よりかなり遅れることもあるので、この措置は完全ではない。

3-3 下部組織の同定

機関レベルより深い組織（大学の学部・大学院研究科や各機関の付属施設など）のレベルで名寄せを行いたい場合はしばしばあるので、機関名辞書では主な機関の下部組織の収録に力を入れている（下部組織収録の考え方については 1-3-2 を参照）。しかし、下部組織レベルの名寄せは機関レベルのそれより一層困難である。その理由は、機関の下部組織表記は、3-1 に述べた表記の揺れが甚だ多様であること、3-2 で述べたように変遷が頻繁であることにもよるが、それ以外に下部組織に特有の別の問題があるからである。

しかし、下部組織レベルの名寄せの必要性が高いので、NISTEP ではそれに意を払っている。ここでは、下部組織の名寄せに特有ないくつかの問題点とそれへの対策を示す。このうち 3-3-1 以外は大学の下部組織に関する問題である。

3-3-1 下部組織名抽出の困難さ

【問題点】

2-1 で述べたように、WoSCC、Scopus においては、代表機関名と下部組織名のフィールド配置は原則としては決まっている。WoSCC では ORG サブフィールドに代表機関名、SUBORG サブフィールドに下部組織名が記入される。Scopus ではもう少し複雑であるが、2-1-4 で述べた置き換え処理をすれば WoSCC に近い形にすることができる。

WoSCC や Scopus の記述がこの原則に沿っていない場合、いろいろな名寄せ上の問題が発生するが、そのことは 3-6 で論ずることにして、ここでは、この原則に沿った場合にも起こり得る問題点を挙げることにする。

(1) 英語正式名のパターンの多様性

下部組織の英語正式名は、次の 3 つのタイプに大別される。これらのいずれの場合にも、代表機関ではなく下部組織に同定される方法を考える必要がある。

(a) 代表機関名の後に下位機関名を続ける

［例］国立研究開発法人国立循環器病研究センター研究所：National Cerebral and

Cardiovascular Center Research Institute

(b) 下位機関名の後に代表機関名を続ける（多くの場合その間にカンマが入る）

［例］公益財団法人結核予防会結核研究所：The Research Institute of Tuberculosis, Japan Anti-Tuberculosis Association

(c) 代表機関名を省略し下位機関名だけで表記する

［例］情報・システム研究機構国立遺伝学研究所：National Institute of Genetics

(2) SUBORG サブフィールドへの所在地データ等の付随

この問題は(1)より遙かに厄介である。SUBORG サブフィールドには、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い。前者の例は"Department of Biology, Graduate School of Science"、後者の例は"Faculty of Engineering, Shinjuku Ku"である。機関名辞書における下部組織の英語名称の大多数は、(1)で述べた(a)または(b)のパターンに属している（すなわち、下部組織名に代表機関名が結合している）ので、このようにそれ以外の単語列が混入していると、SUBORG+ORG マッチあるいは ORG+SUBORG マッチ（2-5-1 参照）を行ってもうまくマッチができない。

【対策】

この問題への対策は、大学とそれ以外の機関との間で大きく異なるので、それぞれについて説明する。

- **大学の下部組織同定への対策**

これについては 2-5-2(1)で述べたとおりである。すなわち、ORG マッチで下部組織を持つ大学に同定されれば、他の機関のように SUBORG+ORG マッチ（あるいは ORG+SUBORG マッチ）を行うのではなく、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う（SUBORG マッチで下部組織の同定がされなければ SUBORG+ORG マッチに進む）。大学に対する WoSCC や Scopus のデータに大学名が省略されることは、皆無ではないが極めて稀なので、このことが可能となる。

ORG マッチで同定された大学の下部組織を SUBORG マッチにより探すので、問題点の(2)を気にする必要はなくなる。

- **大学以外の機関の下部組織同定への対策**

大学以外の下部組織同定にも大学と同様の方法が適用できればよいが、それらの正式名称データには(1)に示した(a), (b), (c)のいろいろなタイプがある。このため、WoSCC や Scopus のデータにも、次のように様々なパターンがある。

① ORG に代表機関名、SUBORG に下部組織名が入っている

② 逆に SUBORG に代表機関名、ORG に下部組織名が入っている

③ ORG に代表機関名と下部組織名が入っている

④ ORG に下部組織名だけが入っている

大学に適用している方法を用いることが難しいため、マッチング方法と機関名辞書の両面で工夫をし

ている。

マッチング方法では、2-5-1 で述べたように、まず ORG マッチ、次いで SUBORG+ ORG マッチを行う。更に 2021 年度から ORG+SUBORG マッチも導入した。これにより、上記①か②のパターンに対しては、正式名が(1)の(a), (b), (c)いずれのパターンであっても下部組織が同定される（代表機関とその下部組織の両方がマッチすれば下部組織を同定機関とする）。しかし、パターン③の場合は(a), (b)のいずれか（ORG 中の並び方による）が代表機関にしか同定されず、パターン④の場合は(a), (b)のいずれもが同定に失敗する。

この解決策として、機関名辞書に Alias または Variant を補う方法がある。(1)の(a)に例示した国立研究開発法人国立循環器病研究センター研究所では、正式名中の代表機関部分と下部組織部分を逆転した Research Institute, National Cerebral and Cardiovascular Center を Alias としている。また、(1)の(b)に例示した公益財団法人結核予防会結核研究所では、正式名中の代表機関部分を省いた Research Institute of Tuberculosis を Variant としている。

しかしながら、問題点(2)に対しては、SUBORG+ORG マッチ及び ORG+SUBORG マッチの適用や機関名辞書への名称補充によりある程度救われるが、十分な解決策を見出すに至っていない。この問題については現在検討を続けている。

3-3-2　第 3 階層以下の組織による表記

【問題点】

1-3-2 で述べたように、機関名辞書では 31 の主要な（発表論文数の点から）大学及び協力を得た 1 大学については第 2 階層の下部組織を網羅的に収録している。しかし、WoSCC や Scopus 中の所属表記においては、第 2 階層下部組織を省略して第 3 階層下部組織が記載されることがしばしばある。つまり、学部名や研究科名を記載せず、その下の学科名や専攻名が記載されているようなケースである。これは情報源の論文に著者がそのように記載しているためである。確かに、Graduate School of Science, XX University よりも Department of Biochemistry, XX University の方が著者の所属や研究分野を具体的に示すので、このような所属表記がなされるのであろうが、これも下部組織同定を難しくする一因になる。大きな大学では、Department of Biochemistry は理学部、医学部、薬学部、農学部のいずれにも存在し得る。学部や学科が安定して存在しておればともかく、大学の組織は始終改組が行われるので、問題は深刻である。

【対策】

この問題への対策は、2-5-2(2)に詳しく述べたとおり、よく現れる第 3 階層組織とその所属先である確率が統計的に高い第 2 階層組織を対応づけた下位機関統計辞書、及びある第 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（多くはその下位の第 3 階層組織）を集めたユーザー定義統計辞書を用いる。これらによって、通常のマッチングでは同定できなかった第 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。

また、よく現れる第 3、第 4 階層組織を機関名辞書に収録することも行っている。

3-3-3　大学に付属する短期大学部

【問題点】

大学に属する短期大学部は機関名辞書では代表機関として扱っているが、WoSCC や Scopus では、大学名を表す部分が ORG サブフィールドに、短期大学部を表す部分が SUBORG サブフィールドに記述されることが多く、そのため短期大学部ではなく大学に同定されてしまう。例えば、大妻女子大学短期大学部(Otsuma Women's University Junior College Division)の"Otsuma Women's University"が ORG サブフィールド、"Junior College Division"が SUBORG サブフィールドに分離していると大妻女子大学に同定されてしまう。

【対策】

このような表記が短期大学部に正しく同定されるように、この種の名称を持つ短期大学部に対し大学と短期大学部のペアを作り、短期大学部を大学の仮想的な下部組織と見なして、大学下部組織の場合と同様の同定手続きを適用することにした(2-5-2(1)参照)。これにより、大学名でマッチした後短期大学部を同定できるようになった。

3-3-4　教員が異なる組織から発表

【問題点】

これは名寄せ上の問題ではなく、同定後に情報分析をする場合の問題である。

一般に大学の教員は学内の複数の組織に所属している。通常、大学には学部(Faculty または School)と大学院研究科(Graduate School)があり、多くの教員はこの両方に所属している。最近ではこの他に、教員が本籍を置く組織を設ける大学が増えている。筑波大学の「系」(Faculty)、金沢大学の「研究院」(Institute)、信州大学の「学域」(Academic Assembly School)などである。これらに重複して所属する教員が論文発表の際どの所属組織を記載するかは、人により、あるいは同じ人でも時期によって様々と思われる。論文発表の際に記載する組織が一定していないことは、組織別の業績を集計・分析する場合に困難をもたらす。

【対策】

NISTEP の名寄せでは特に対策はとっていない。名寄せ後の集計・分析の際には、その目的に応じて合体や按分を行うことになろう。対象が論文情報の場合は、著者名と組み合わせることも考えられる。

3-4　同一の名称を持つ異なる機関

異なる機関の英語名が同一であると、当然のことながらどちらに同定するか難しい。以下、次の 3 つの場合に分けて問題点と対策を論ずる。

(1) 変遷の前後で日本語名は変わるが英語名は変わらない場合

(2) かつて存在した機関あるいは組織の英語名を、新しくできた同系列の機関あるいは同機関内の組織が再度用いる場合

(3) 無関係の機関で日本語名も異なるが、英語名の綴りがたまたま一致する場合

3-4-1　変遷前後の機関

この問題については 3-2 で例示し、対策も述べたので参照されたい。

3-4-2　直接の継承関係にないが関係のある機関・組織

【問題点】

英語名が同一のこのような関係には次の 3 タイプがある。

(a) 一旦他の名称の機関になったが再び過去の名称に復帰

代表的なのは東京都立大学→首都大学東京→東京都立大学の例である。この場合、日本語正式名、英語正式名(Tokyo Metropolitan University)とも新旧同じである（首都大学東京も同じ英語名）。民間企業にも、シチズン時計株式会社(Citizen Watch Co., Ltd.)→シチズンホールディングス株式会社(Citizen Holdings Co., Ltd.)→シチズン時計株式会社(Citizen Watch Co., Ltd.)など同様の例がいくつかある。

(b) 同一系列内の過去存在機関と同一名の新機関が出現

(a)と違って、同一英語名称の機関が直列的な関係にはない場合である。（旧）ダウ・ケミカル日本株式会社(Dow Chemical Japan Ltd.)は別の名称の会社に変更されたが、その後暫く経ってから同じ系列に同名の会社（英語名称も同一）が設立されたのが、今のところ唯一の例である。

(c) 同一機関内の過去存在組織と同一英語名の新組織が出現

(b)と似ているが、これは同一機関内の組織であり、日本語名称は互いに異なる。

例としては、金沢大学の医学部と医薬保健研究域医学系（英語名はいずれも Faculty of Medicine, Kanazawa University）及び薬学部と医薬保健研究域薬学系（英語名はいずれも Faculty of Pharmaceutical Sciences, Kanazawa University）がある。2008 年に医学部、薬学部が廃止されたが、その後にできた学部組織は金沢大学医薬保健学域とその下の各学類からなり、医薬保健研究域は教員組織であるため、「学部」と「研究域」の間に継承関係を付けることができない。

別の例として、東京工業大学の理学部と理学院（英語名はいずれも School of Science, Tokyo Institute of Technology）、工学部と工学院（英語名はいずれも School of Engineering, Tokyo Institute of Technology）がある。この大学では 2016 年度に大幅な組織の改編が行われ、学部(School of XXX)と大学院研究科(Graduate School of XXX)を統合した院(School of XXX)が作られた。学部と院の英語名が同じ School of ～で始まるので、School of Science と School of Engineering はどちらを指すのか、その情報だけでは識別できない。院は学部と大学院が合体した組織であること、改組に当たって学部と院の間で学科の移行が交差していることから、学部と院の間に継承関係を設定することは不適当であり、継承関係を用いる識別はできない。

以上、(a), (b), (c)いずれの場合も、同一の英語名を持ちながら直接の継承関係にないので、3-4-1 の場合と同じ対策を採ることができない。（但し、東京都立大学の場合は、間に挟まれた首都大学東京も同

一の英語名なので、3-4-1 の方法を用いてどの時期の大学かを識別することができる。）

【対策】

英語名だけでなく日本語正式名も同一の機関（上記の(a)と(b)）については、機関名辞書への登録をどうするかの問題があるが、これについて 1-2-2 で述べた。

(a), (b), (c)に共通する名寄せの際の対処として、機関名辞書の英語名称は同一だがそのうち片方の機関が非現存になっている場合に、論文の発表年が旧機関の廃止（統合、名称変更を含む）の年より新しければ旧機関への同定を自動的に阻止する特別ルールを導入した。

しかし、機関名辞書には機関の創設年の情報がないため、発表年が旧機関の廃止年以前だと、新旧両機関の同時同定となる。この問題があるため、特に出現回数が多い東京工業大学の理学部と理学院、及び工学部と工学院については特別ルールを導入した。東京工業大学(Tokyo Institute of Technology)が同定された後、School of Science（または School of Engineering）の単語列があれば、論文発表年が 2016 年以前なら理学部（または工学部）に、2017 年以降なら理学院（または工学院）に同定するというものである。

3-4-3 関係のない機関

【問題点】

無関係の機関の英語名が一致するのは、日本語名の漢字は異なるがローマ字綴りで同一になる場合である。医療機関の英語名称には単に XXX Hospital あるいは XXX Medical Center とするところが多いので、この例が屡々生じる。医療法人医仁会武田総合病院と一般財団法人竹田健康財団竹田綜合病院（ともに Takeda General Hospital）、医療法人財団順和会山王病院と医療法人翠明会山王病院（ともに Sanno Hospital）等の例が見られる。医療機関以外では聖泉大学と清泉女子大学（ともに Seisen University）、TOA 株式会社と東亜建設工業株式会社（ともに TOA CORPORATION）等の例がある。

【対策】

次の 2 つの方法のいずれかを適用している。

まず、一方の機関が WoSCC や Scopus にほとんど出現しない場合は、その機関の名称を NotUse にする（（1-2-3(4)参照））。新菱冷熱工業株式会社と株式会社新菱の英語名はどちらも SHINRYO CORPORATION であるが、後者の名称を NotUse にしている。

両方の機関がある程度出現する場合は、2-5-3(2)で述べた特別ルールを設定する。問題点で例示した 4 組にはこの方法を適用している。いずれも 2 つの機関の所在地が異なることを利用し、所在地の地名や郵便番号、及び機関の特徴を表す語によりどちらが適切な同定機関かを判別する。聖泉大学と清泉女子大学での識別方法を 2-5-3(2)に示したが、他についても似た方法による。なお、Sanno Hospital の場合は、この名称の 2 機関に医療法人社団高邦会福岡山王病院(Fukuoka Sanno Hospital)を含めた 3 機関で特別ルールによる識別を行う。

同定対象データ中に特徴を表す文字列が見当たらない場合の処置はルールにより異なる。聖泉大学／清泉女子大学の場合及び TOA 株式会社／東亜建設工業株式会社の場合は 2 機関の同時同定とするが、医療機関同士の場合（例示した 2 組以外も含む）は同定機関なしとする。同定なしとする理由は、医療機関の場合、機関名辞書に登録していないが同一の単語列を含む英語名を持つ病院があり、それらが正解である可能性があるためである。

3-5　類似の名称を持つ異なる機関

3-5-1　似た名称の大学

この問題については 2-5-3(1)で例示し、その主要な対策が特別措置機関統計辞書を用いた誤同定の防止であることもそこで述べた。しかし、この方法では十分に対応できない 2 つの事例があり、それらについては特別ルールの設定により対処している。

そのひとつは、東京理科大学とそれを母体とする大学の間の識別である。特別措置機関統計辞書ではなく特別ルールを適用したのは、次のような複雑性があるためである。

(a) 関係する大学には、東京理科大学、山口東京理科大学と 2016 年にそれを継承した山陽小野田市立山口東京理科大学、諏訪東京理科大学と 2018 年にそれを継承した公立諏訪東京理科大学の 5 つがある。公立大学となった後の英語名には"Tokyo University of Science"が含まれていないが、その後も旧名が使われていることがある。特別ルールではこれらの点も考慮して同定を行う。

(b) 東京理科大学の英語名は Tokyo University of Science であるが、Science University of Tokyo の所属表記もよく使われる。

(c) WoSCC や Scopus の ORG サブフィールドに記入された"Tokyo Univ(ersity) Sci(ence)"が、東京大学の理学部あるいは理学系研究科を指す揺らぎ表記であることがときどきある。特別ルールでは、所在地単語列中に"Bunkyo","Hongo"あるいは"1130033"等があれば東京理科大学を同定候補から除くとしている。

もうひとつは、浜松医科大学と浜松大学（2013 年に統合されて常葉大学になる）の識別である。英語名はそれぞれ Hamamatsu University School of Medicine と Hamamatsu University であるが、前者が ORG サブフィールドに"Hamamatsu University"、SUBORG サブフィールドに"School of Medicine"と分離してしまうことがよくあり、同定の混乱が起こる。当初は特別措置機関統計辞書により、医学関係の語句が共出すれば浜松医科大学、"Faculty of Administration"等が共出すれば浜松大学として判別していた。しかし、浜松医科大学の下部組織である医学部附属病院も機関名辞書に登録しているため、この方法では解決が難しい問題があることが判り、特別ルールに切り替えた。

3-5-2　ありふれた語からなる表記

【問題点】

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して（その

中には機関名辞書に登録していない機関もあり得る）同定の混乱が起こりやすい。2-5-3(3)にいくつかの事例を示した。ありふれた語ではないが、民間企業の系列やグループの中では、語の一部が一致し、そのために同時同定や誤同定が起こることがある。また、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（一般財団法人科学技術振興会の略称 FAST など）も誤同定を引き起こす可能性が高いが、この件については 3-1-3 で述べた。

【対策】

2-5-3(3)と 3-1-3 でも対策を述べているが、まとめると以下のようになる。

(1) 機関名辞書への登録

類似の名称を持つ機関（下部組織を含む）を登録し、それらに使われるいろいろな表記を Variant として収録する。機関名辞書の登録基準から外れた組織を、このために登録することもある。

(2) 一部名称の NotUse 指定

問題となる機関英語名が WoSCC や Scopus にほとんど出現しなければその英語名を NotUse にする。次の例がある（いずれも後者の名称を NotUse にしている）。

- 株式会社クボタ(KUBOTA CORPORATION)と株式会社クボタ商会(KUBOTA CO., LTD.)
- 株式会社クレハ(KUREHA CORPORATION)と呉羽テック株式会社(KUREHA LTD.)
- 三共株式会社と株式会社三共（どちらも SANKYO CO., LTD.）

(3) 特別ルールの設定

以上の方法で対処が難しい場合は特別ルールを設ける。いくつかの例を挙げる。

- 公益財団法人環境科学技術研究所（2012 年以前は財団法人）

多くの機関が、この機関の英語名 Institute for Environmental Sciences を包含する名称を持つ。これに対する特別ルールは 2-5-3(2)に示した。

- 旧工業技術院の機械技術研究所

工業技術院の研究所は 2001 年に廃止されて独立行政法人産業技術総合研究所（現在は国立研究開発法人）に統合されたが、それ以前の WoSCC や Scopus には数多く現れる。機械技術研究所はそのひとつであるが、その英語名 Mechanical Engineering Laboratory は、いくつかの他の機関の附属研究所と同一である。特に出現頻度が多いのは、株式会社日立製作所の機械研究所と株式会社神戸製鋼所の機械研究所である（両者とも正式の名称は Mechanical Engineering Research Laboratory であるが Mechanical Engineering Laboratory と表記される場合が多い）。これらの研究所は機関名辞書に登録されていないので、特別ルールでは、Mechanical Engineering Laboratory が Hitachi と共出するときは株式会社日立製作所に、Kobe と共出するときは株式会社神戸製鋼所に、そしてこれらの共出がないときは工業技術院機械技術研究所に同定する。

- 独立行政法人国立病院機構榊原病院と社会医療法人社団十全会心臓病センター榊原病院

英語正式名はそれぞれ Sakakibara National Hospital、Sakakibara Heart Institute of Okayama であるが、前者は NHO Sakakibara Hospital、後者は Sakakibara Hospital

Cardiovascular Center 等の Variant で表記されることも多く、混同されやすい。そこで、Sakakibara Hospital にマッチした場合、Cardiovascular あるいは Okayama と共出すれば国立病院機構榊原病院を同定候補から外し、これらのいずれの語とも共出しなければ国立病院機構榊原病院に同定するという特別ルールを設定した。

3-5-3　所管・共同運営等の関係がある機関

【問題点】

同定対象データ中の国立試験研究機関あるいは独立行政法人の英語名称に所管する省庁名が含まれていると、省庁と研究所の両方に同定される。また、県の公設研究機関の名称中に"XXX Prefectural Government"が含まれていると、その県庁にも同定されてしまう。

次に、共同運営や事業協力の関係にある機関が、名称中に似通った単語列を共有するため同定が混乱する例を 2 つ挙げる。

(a) SPring-8

SPring-8 は播磨科学公園都市にある大型放射光施設で、運営を国立研究開発法人理化学研究所放射光科学研究センター、利用促進業務を公益財団法人高輝度光科学研究センターが担っている。また、実際の運用に関する諸業務を担当するスプリングエイトサービス株式会社もある。これらの機関や組織の英語名には"SPring-8"が含まれるが、この施設は国内外に広く開放されているので、多くの機関が発表する論文の所属表記にも"SPring-8"の文字が含まれる。従って、所属機関データに"SPring-8"が含まれるとき("SPring 8", "SPring8"などと表記されることもある)、どの機関に同定すべきか判別する必要がある。

(b) 一般財団法人阪大微生物病研究会(2010 年以前は財団法人)

この団体が大阪帝国大学(当時)の提携機関(今で言う大学発ベンチャー)として 1934 年に発足したことからこの名が付けられた。英語名も The Research Foundation for Microbial Diseases of Osaka University で大阪大学微生物病研究所(Research Institute for Microbial Diseases, Osaka University)とよく似ているため同定に混乱が生じる。

【対策】

国立研究所(あるいは独立行政法人)の機関表記に所管の省庁名が含まれたり、県の公設研究機関に県庁名が含まれたりすることによる同時同定の問題は、機関名辞書において、対となる機関の間に所管-被所管の関係づけを行うことにより解決を図っている(1-2-7(2)を参照)。この関係が付けられた対の両機関にマッチすると、所管される方の機関に同定がなされる(2-5-1 参照)。

次に、共同運営や事業協力の関係にある機関の同時同定の問題には、特別ルールの設定により対処している。上に挙げた例の(a)については、機関名に"SPring-8"を含む 3 機関の他、SPring-8 を利用した研究をよく発表する 4 機関を含めて、それぞれの機関に特有の単語列との共起により、いずれに同定するかを判別する(いずれに該当する単語列もない場合は、国立研究開発法人(または独立行政法人)理化学研究所に同定)。例(b)については、単語列中に"Research Foundation Microbial Disease"を

含めば阪大微生物病研究会に、そうでなければ大阪大学のいずれかの組織に同定する。

3-6　WoSCC、Scopus における原則から外れた表記

2-1 で述べたように、通例は ORG サブフィールドに代表機関、SUBORG サブフィールドに下部組織が記述されているが、この原則から外れた表記も見られる。特に Scopus は不規則性が高い。以下にはこれをいくつかのパターンに分けて説明し、機関名辞書あるいは名寄せプログラムで用いている対策を示す。

3-6-1　ORG と SUBORG の逆転

【問題点】

原則とは逆に、ORG サブフィールドに下部組織名が、SUBORG サブフィールドに代表機関名が入っている場合がある。

【対策】

まず、SUBORG に"University"または"Univ"の語があれば ORG と SUBORG を交換したマッチングを行う。これにより大学の場合は通常通りの同定処理がなされる。それ以外の場合、ORG マッチでは機関が同定されないが、それに続く SUBORG+ORG マッチあるいは ORG+SUBORG マッチで多くの場合下部組織の同定がなされる。少なくとも代表機関には同定される。しかし、これらにより完全に解決されるわけではない（3-3-1 の問題点(2)で述べた SUBORG サブフィールドに所在地データが混入している場合など）。

3-6-2　ORG に代表機関と下部組織が合体した表記

【問題点】

この問題は大学において生じる。ORG で下部組織を持つ大学が同定された場合は、SUBORG でその大学の下部組織とのマッチを行うためである（2-5-2(1)を参照）。

東京大学医学系研究科(Graduate School of Medicine, the University of Tokyo)を例に説明する。原則に従えば ORG サブフィールドに"The University of Tokyo"、SUBORG サブフィールドに"Graduate School of Medicine"が表示されるが、ORG サブフィールドに[A]"Graduate School of Medicine, the University of Tokyo"あるいは[B]"The University of Tokyo Graduate School of Medicine"のように合体して記載されている場合がある。

大学附属の病院、博物館、図書館等では、ORG サブフィールドに"XXX University Hospital"、"XXX University Museum"、"XXX University Library"などと記載されることが多い。これらの組織では正式名に"XXX University"が冠せられることが多いことによる。しかし、機関名辞書の正式名は、その命名原則により、"XXX University Hospital, XXX University"である。

【対策】

上述の東京大学医学系研究科の場合、表記[A]は機関名辞書の正式名にマッチするので正しく同定

されるが、表記[B]はマッチせず、代表機関である東京大学に同定される。このような場合に対処するため、機関名辞書に下部組織情報を持つ大学に限り、機関名辞書通常の語順に加え、大学名と下部組織名を反転させた語順でもマッチングを行うようにした。

大学附属病院等に対する上記の表示の場合、ORG マッチにより代表機関の XXX University は同定されるが、SUBORG サブフィールドに下部組織名がないので、通常の方法では下部組織の病院等までにはマッチしない。そこでこのような組織には、正式名の"XXX University Hospital, XXX University"の他、Variant に"XXX University Hospital"等を追加することにより、ORG マッチでの下部組織同定を可能にした。

3-6-3　ORG に下部組織の一部が混入

【問題点】

この問題はかなり厄介である。事例の多い東京大学を例に説明する。

東京大学の英語正式名は The University of Tokyo であるが、Tokyo University と表記されることもしばしばある。ORG サブフィールドの"Tokyo Univ"の後に学部あるいは大学院研究科の名称の一部が紛れ込んだ表記がなされると、Tokyo University of ...という英語名を持ついろいろな大学と混同しやすい。"Tokyo Univ Agr"は東京農業大学(Tokyo University of Agriculture)か東京大学農学部か、"Tokyo Univ Pharm"は東京薬科大学(Tokyo University of Pharmacy and Life Science)か東京大学薬学部か、"Tokyo Univ Sci"は東京理科大学(Tokyo University of Science)か東京大学理学部か、などである。

このような例は他にも多数挙げることができる。"Okayama Univ Sci"は岡山理科大学(Okayama University of Science)と思われるが、岡山大学理学部（Okayama University, Faculty of Science）である可能性も否定できない。"Hokkaido Univ Educ"は北海道教育大学(Hokkaido University of Education)と北海道大学教育学部の可能性がある。

【対策】

この問題については特段の対応をしていない。なぜなら、対応しなければ多くは単科大学（～農業大学、～薬科大学、～理科大学、～教育大学等）に同定され、それが正しい確率が高いからである。総合大学の学部をこのような形式で表記するのは省略のし過ぎであり、同定漏れになってもやむを得ないと考えられる。東京理科大学と東京大学理学部の場合だけは、東京理科大学系統の 3 大学識別のための特別ルールを作る際に考慮したが（3-5-1 参照）、これは例外的である。

3-6-4　代表機関名が ORG と SUBORG に分離

【問題点】

代表機関名が長く、見かけ上 2 つの部分に分けられる場合がある。このようなとき、その代表機関名が ORG サブフィールドと SUBORG サブフィールドに分離してしまうことが屡々起こる。"XXX University School of Medicine"の名を持つ医科大学では、"XXX University"が ORG に、"School of Medicine"

が SUBORG に分離する。浜松医科大学(Hamamatsu University School of Medicine)、聖マリアンナ医科大学 (St. Marianna University School of Medicine)がそうであるが、獨協医科大学(Dokkyo Medical University)も"Dokkyo University School of Medicine"とよく表記されるので同様なことが起こる。そうすると、当該の大学には同定されないばかりか、浜松医科大学の場合は浜松大学(Hamamatsu University)に、獨協医科大学の場合は獨協大学(Dokkyo University)に同定されてしまう。

ほとんどの国立高等専門学校の英語名は National Institute of Technology, XXX College であるが、これが屡々"National Institute of Technology"と"XXX College"に分離する。"National Institute of Technology"という名の機関はないので、このままだと同定されない(あるいは、たまたま"XXX College"という名の短期大学あるいは大学があるとそれに誤同定される)。

【対策】

上述の浜松医科大学の場合は、浜松大学との識別のための特別ルールを設けている(3-5-1 参照)。また獨協医科大学と獨協大学は、特別措置機関統計辞書の中のペアのひとつである(2-5-3(1)参照)。これらによって、"School of Medicine"がアドレスフィールド中に存在すれば浜松医科大学または獨協医科大学に同定される。

国立高専の問題については、独立行政法人国立高等専門学校機構(英語正式名 National Institute of Technology, Japan)の Alias に"National Institute of Technology"を加え、この名称で始まるすべての国立高専をその下部組織と見なして、大学と同じ同定法を適用することにした。

しかし、代表機関名が ORG と SUBORG に分離して同定がうまくいかない問題の全面的解決はなかなか難しい。地方自治体の公設研究機関には長い英語名を持つものが多いので同様の問題が起きている可能性があるが、対応できていないケースがあると推測される。

3-6-5 主要な名称情報が ORG と SUBORG から欠落

【問題点】

これは Scopus における問題点である。ORG 及び SUBORG サブフィールドに必要な情報がほとんどなく、ADDRESS サブフィールドに詳細な機関表記が含まれている場合がある。

【対策】

これについては今のところ対策が採られていない。一時、Scopus データにおける ADDRESS サブフィールドの文字数が一定数以上の場合、このサブフィールドに対して同定を実施することにしたが、この方法にも問題が起きることが判り取りやめた。

3-7 複数機関の同時同定

同定処理が一応終わった段階で、複数の機関が同定候補として残ることがある。しかし、同定さるべき機関は実はそのうちの 1 つだけであることがかなりの割合で存在する。これを選び出すため、2-5-4 で述

べた絞り込みを行う。

絞り込みを行ってもなお 2 つ以上の機関が残った場合、現在の絞り込み法では発見できない重複同定である可能性もあるが、それを除けば、同時同定が正しいと考えられる。すなわち、その所属機関データに該当する著者は複数の機関に所属している。この詳細については 2-1-2 に述べた。

このようなことがあるために、2-5-1 で述べたように、一旦同定機関が見つかっても、同定部分を除いた単語列に対して再び同定処理を行うこと（再帰同定）が必要である。

所属機関データに 2 つの機関が存在するとき、両方の機関が ORG サブフィールドに入っている場合と、一方が ORG サブフィールドに、他方が SUBORG サブフィールドに入っている場合がある。

4　NISTEP 名寄せプログラムの性能、及び名寄せ結果の調査・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（同定フラグが"S", "H", "N"のいずれか）が現れる。また、第 2 章と第 3 章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。このため、一定以上の出現頻度があったアドレスフィールドデータに対し、目視によって同定失敗の理由を考察し、対策を検討している。

この章では、NISTEP で行っている名寄せの性能を報告するとともに、データチェックの方法、性能の改善のために対処する方法を示す。

4-1　名寄せプログラムの性能

名寄せプログラムの性能は、次の 2 つの面から測ることができる。

(i) 充足率：同定対象レコードのうち何パーセントが辞書中の機関に同定されるか。逆に言うと何パーセントのレコードが機関同定できないか。

(ii) 正解率：機関同定されたレコード中何パーセントが正しく同定されているか。逆に言うと誤同定は何パーセントか。

ここでは、最近実行した WoSCC の著者所属機関データの名寄せ結果に基づき、これらについて記す。Scopus は WoSCC に比べて充足率、正解率ともやや低いが、傾向はほぼ一致している。

4-1-1　名寄せの充足率

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。同定フラグ別では、93.7%が最長マッチ（同定フラグ"L"）で同定された。特別ルールによる同定（2-5-3(2)参照、同定フラグ"R"）が 0.8%、混合マッチ（同定フラグ"M"）、ベクトルマッチ（同定フラグ"V"）は合わせて 0.05%と僅少であった。残りの 5.4%は機関の同定ができなかった（同定フラグ"S", "H"または"N"）が、セクターを同定できた"S"を加えると、充足率は 96.2%になる。

Scopus に対する充足率はやや低く 92.5%であるが、内訳はほぼ同様の結果になる。

4-1-2　名寄せの正解率と誤同定のタイプ

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、通常は次のようにしている。

(1) 機関同定されたレコード（機関同定ができなかったレコードについては 4-2 で述べる）を次の 3 つの集合に分ける。

・集合Ⅰ：単一の代表機関に同定されたレコード

・集合Ⅱ：単一の下部組織に同定されたレコード

・集合Ⅲ：複数の機関が同時同定されたレコード（集合Ⅰ、Ⅱに比べてごく小さい）

(2) 同定された機関（Ⅲの場合は同定された機関の組）ごとにレコードをまとめ、次に同じ同定機関（あるいは同定機関の組）の中で機関データ（2-1-1(1)で述べた ORG, SUBORG, ADDRESS サブフィールド）が全く同じレコードをまとめる。

(3) 同定機関と機関データが同じレコードの重複を除き 1 個だけを残す。このとき、重複するレコードの数（出現頻度に当たる）を記録しておく。これによってレコード数は当初の 1/5 程度になる。

(4) ある出現頻度以上のレコードをチェックの対象とする。但し、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ってしまうので、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。

(5) 目視でチェックを行う。同定が誤っているデータには以下の情報を追記する。

(a) 正解の機関：辞書に登録されていない機関が正解の場合や、正解機関が不明の場合もある。

(b) 誤りのタイプ：後述する（図表 10 を参照）。

(c) 誤同定の理由

最新の名寄せの調査が未了なので、ここでは、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結果を示す。上記の 3 つの集合別のレコード数は図表 9 の通りである。

図表 9　全同定レコード数とチェック対象レコード数（出現頻度の合計）

同定レコードの集合	全同定レコード数	チェックレコード数	チェックレコード比率
Ⅰ.単独同定（代表機関）	2,381,914	807,516	33.9%
Ⅱ.単独同定（下部組織）	2,167,623	807,516	45.8%
Ⅲ.複数機関同定	51,546	18,819	36.5%

チェック対象は全レコードの 1/3～1/2 程度であるが、ORG 及び SUBORG は共通で ADDRESS のみが異なる（主に所在地データや郵便番号の有無、表記の違いによる）レコードがチェック対象外に多数あるので、実際には 7～8 割のレコードをチェックしたものと推定される。

図表 9 のチェック対象レコードを調査した結果を図表 10 に示す。表中斜線を引いたセルはあり得ないタイプで、"0"はあり得るが存在しなかったタイプである。

単独同定（代表機関）（集合Ⅰ）と単独同定（下部組織）（集合Ⅱ）の正解(O)の率はそれぞれ 98.4%、96.0%（つまりエラー率は 1.6%、4.0%）であるが、代表機関の誤り(A1+A2)はそれぞれ 0.02%、0%と極めて低い。集合Ⅰでは代表機関に同定されたがその下部組織が正解であるもの(B1)が 1.3%、集合Ⅱでは変遷前または変遷後の機関が正解であるもの(C)が 4.0%で、エラーの大部分またはすべてを占める。一方複数同定（集合Ⅲ）は過半(53.9%)がエラーであるが、その大部分(50.6%)は、同一代表機関の機関内の 2 つの下部組織に同時同定されたが一方だけに同定すべきもの(B3)である。これらの誤同定の内容については 4-3 で説明する。

図表 10　2020 年実施の WoSCC 名寄せのチェック結果

同定の正誤と誤りのタイプ	Ⅰ.単独同定（代表機関）	Ⅱ.単独同定（下部組織）	Ⅲ.複数同定
O:正しい同定	794,638	952,276	8,683
A1:代表機関の誤り（正解機関が辞書にあり）	173	0	0
A2:代表機関の誤り（正解機関が辞書になし）	0	0	0
A3:複数同定の一方の代表機関は不要（単独同定が正しい）			480
B1:代表機関同定だがその下部組織が正解	10,881		0
B2:同じ代表機関の別の下部組織が正解		0	45
B3:複数同定の一方の下部組織は不要（単独同定が正しい）			9,524
C:変遷前または変遷後の機関が正解	1,824	39,563	87
チェックレコード数	807,516	991,839	18,819

3 つの集合Ⅰ、Ⅱ、Ⅲではそれぞれチェックしたレコードの抽出率が異なるので、図表 10 に示した数字からそのまま全集合の分布を求めることはできない。図表 9 に示す抽出率に基づいて、全集合における正解とエラーの分布を推定した結果を図表 11 に示す。このとき、チェックしたデータにおける分布が全体に適用されると仮定した。

図表 11　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 [4]や Scopus-NISTEP 大学・公的機関名辞書対応テーブル [5]のエラー率はずっと低い。

4-2　機関同定できなかったデータの調査

4-1-1 で述べたように、WoSCC や Scopus の名寄せでは、全体の 5～7%のレコードが機関同定できないので、20 年以上のデータを処理すれば、20～30 万もの未同定データが生ずる。通常はこのうちから、同一の ORG+SUBORG+ADDRESS 情報が 15～20 レコード以上のデータを抽出して、未同定となった理由を検討する。これらの中には、代表機関名がなく下部組織名のみが記されたもの（例えば、大学名がなくて"Graduate School of Engineering"以下の情報のみのものなど）、表記が自己流であった

り省略が過度であったりするため機関の判別ができないものも多いが、機関が判別できるものについては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されている。このような例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等にはこの選択肢を採る。また、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も採択を見送る。

補足すると、機関の登録や Variant の追加により充足率を上昇させることは、表記の識別の難しさによる正解率の低下を伴いがちである。つまり両者はトレードオフ関係にある。第 2 章の冒頭で述べたように、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。従って、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ので、現在の充足率(93～95%)はやむを得ないと考えている。

4-3　主な誤同定の内容とそれへの対処

4-1-2 のチェック作業で誤同定と判定されたデータについて、作業に入る前の元データファイルに戻って個々のレコードを修正する。正解機関が辞書未登録機関や不明の場合は、同定フラグを"N"にする。

図表 11 で実際にエラーが存在した A1、A3、B1、B3、及び C について、その主な内容と対処の方法を示す。

(1) A1:代表機関の誤り（正解機関が辞書にあり）

このタイプの誤同定はごく少数であるが、次のような事例があった。

- 大学の短期大学部が大学に同定された。これに対しては、2-5-2(1)、3-3-3 に述べたように、「大学・短期大学ペア定義テーブル」を作ることにより短期大学部に同定されるようにした。
- 浜松医科大学の Hamamatsu Univ, Sch Med が ORG と SUBORG に分離したため浜松大学に同定された。3-5-1 に述べたように、特別措置機関統計辞書による識別を特別ルールによる識別に変更することにより解決した。

(2) A3: 複数同定の一方の代表機関は不要（単独同定が正しい）

この多くは、ある機関の Variant が他の機関にマッチしたものであり、その Variant を削除あるいは修正することにより解決した。たとえば、ORG="Nagasaki Med Ctr", SUBORG="Natl Hosp Org"は国立病院機構長崎医療センターが正解だが、SUBORG+ORG マッチで国立病院機構長崎病院の Variant である"Natl Hosp Org Nagasaki"にもマッチしてしまう。そのため、このような例が起こり得る他の国立病院機構の病院も含めて、Variant の"Natl Hosp Org XXX"または"NHO XXX"を削除した。

この他、地方自治体の機関の名称に"XX Prefectural Govt"が含まれていたため県庁に同定されたものがあった。1-2-7(2)、3-5-3 に述べたように、この種の機関の間に所管-被所管の関係づけを行うことにより解決した。

(3) B1: 代表機関同定だがその下部組織が正解

このタイプのエラーは、代表機関同定のエラーの約 8 割を占める。中でも大きな割合を示すのは、下部組織を包括的に収録する 32 大学(1-3-2(1)参照)に関するものである。これについては特に注意を払ってチェックし、Variant の補充、新設組織の辞書への登録等により対処している。

それ以外では、ORG="Natl Canc Ctr", SUBORG="Res Inst"が国立がん研究センター研究所(National Cancer Center Research Institute)に同定されず上位の国立がん研究センターに同定されるような例が多く見られた。これに対しては、3-3-1 で述べたように、SUBORG+ORG マッチに加えて ORG+SUBORG マッチの導入、及び Alias または Variant の追加(上の例であれば"Research Institute, National Cancer Center"を Alias とする)が対処法になるが、SUBORG サブフィールドへの所在地データ等の混入の問題があって十分な解決には至っていない(3-3-1 参照)。

なお、誤同定ではないが、1-3-2 に示した基準には入っていないため辞書に未登録の下部組織から、多数の論文が WoSCC や Scopus に収録されていることがある。これらはその代表機関に同定されるので、そのような下部組織を見出して辞書に登録することがある。

(4) B3: 複数同定の一方の下部組織は不要(単独同定が正しい)

複数同定のエラーではこれが大部分を占めていた。そのほとんどが 32 大学に関するもので、重要なのは以下の 2 タイプである。

- 下位機関統計辞書またはユーザー定義辞書による同定追加

  32 大学の下部組織同定では、SUBORG または ADDRESS で最長マッチがなされれば下位機関統計辞書またはユーザー定義統計辞書による同定は行わないとしている(2-5-2(3)を参照)が、ORG に"XX Univ Hosp"等があって直接下部組織に同定された場合は、この処置がなされていなかった。このため、附属病院がある大学で病院と他の下部組織(多くは医系の学部または医学研究科)の同時同定が多数起こった。例えば、ORG に"XX Univ Hosp"があったため大学附属病院にマッチし、下位機関統計辞書で"Dept Surgery"がその大学の医学研究科に関連付けられているような場合である。ORG マッチで下部組織同定された場合には下位機関統計辞書マッチは行わないようにプログラムを修正することにより、この問題はほぼ解決した。

- 継承関係のない同名機関の同時同定

  直接の継承関係がないが英語名が同一の機関の間では、継承関係を用いる識別ができず、同時同定が起こる。この問題については、事例、対処法を含めて 3-4-2 に詳しく述べたのでここでは繰り返さない。そこでは、この種の関係に対する特別ルールを設定したこと、しかしこのルールで完全には解決されないことを述べた。

  変遷が度々行われている機関(主に国立研究開発法人)に対しては、2-5-4(5)で述べた 2 機関

比較法も適用している。

(5) C: 変遷前または変遷後の機関が正解

図表 11 に示すとおり、このタイプのエラーが最も多い。英語機関名が同一または類似の変遷前後の機関間の識別方法は、2-5-4(4)及び 3-2 で述べたとおりであるが、それでも両機関を取り違えた同定が発生する。

この原因のほとんどは次の 2 つに帰される。

(i) 継承前から存在している下部組織が一方にのみ登録されている。代表機関 A が B に継承され、a を A の下部組織、b を B の下部組織としたとき、b のみが辞書に登録され、a は登録されていない場合、a と b の英語名が同一あるいは極めて類似していると、その英語名表記のデータが a を指していたとしても b に同定されてしまう。

(ii) a も b も辞書に登録されており、それらの英語正式名 name-a と name-b が互いに異なるとする。しかし、変遷後でも旧名である name-a の名称をそのまま使った発表があると(実際屡々存在する)、name-a が b の Alias または Variant として辞書に収録されていなければ a に同定されてしまう。同じことは、代表機関である A と B の間でも起こり得る。

(i)に対しては未登録の下部組織 a(b の場合もあり)を辞書に登録することにより、(ii)の場合は B または b の Alias または Variant に継承期間前の名称を追加することにより対処している。

以上に述べたようなチェックと対処により、誤同定は大幅に減っていると推測される。図表 10、図表 11 の数字は 2000 年に行った WoSCC 名寄せに基づくが、その後、比較的誤同定が多かった B1、B3、C 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　おわりに－未解決の課題

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。この報告書の最後に、代表的な今後の課題を述べることとする。

(1) 下部組織を収録する大学の拡張

1-3-2(1)で述べたように、下部組織を包括的に収録している大学は32大学に過ぎない。NISTEPの種々の調査分析では、生産する全論文数に対するシェアにより大学を第1～第4グループ及びその他に分けている[15]が、上記32大学は、第1グループと第2グループのすべて（それぞれ4大学、14大学）、第3グループの25大学中13大学、第4グループの136大学中1大学である[8]。第3グループと第4グループの残りの大学の下部組織を順次収録して行くことが目標である。

(2) 名寄せプログラムの利用者拡大のための方策検討

名寄せプログラムの利用者や利用希望者からは、次のような要望がある。これらの要望に対しては、順次対応を考えている。

- WoSCC、Scopus以外の情報源への対応：特に、現在未対応である日本語の機関名の名寄せは、科研費データ、Researchmap等の分析に重要である。英語のデータに対しては現在にプログラムで処理可能であるが、WoSCC、Scopus以外の主要なデータ源に対する同定性能の改善へのニーズが強ければ対応したい。
- 機関名辞書への機関やデータの追加：利用者が、独自の必要性により機関名辞書にない機関やデータを追加することは現在でも可能だが、その手順がやや複雑である。特に、自機関の登録されていない下部組織の追加が望まれている。
- WoSCC、Scopusのデータの入力ファイルへの取り込み：これらのデータベースの利用者の多くは、オンラインデータベースからのデータ（サブフィールド分割されていない）を用いているので、それを名寄せする場合、入力ファイルのORG、SUBORG等のサブフィールドに簡単に分割することが求められている。

(3) 機関名辞書への創設年の導入

[8] 参考文献15では、第3グループは26大学、第4グループは137大学であるが、その後第3グループの2大学（大阪府立大学と大阪市立大学）と第4グループの2大学（大阪医科大学と大阪薬科大学）が統合し、それぞれ大阪公立大学、大阪医科薬科大学になったので、ここでは第3グループ25大学、第4グループ136大学とした。

機関名辞書には、機関が非現存になった年の情報はあるが、いつ創設されたかの情報はない。従って、名寄せ時に変遷年による機関の識別が不完全であり、創設される前のデータがその機関に同定されることがある(3-4-2 に挙げた例を参照)。しかし、機関名辞書の全機関について創設年を調査するには膨大な作業が必要であり、見通しが立っていない。

(4) 名寄せプログラムの一層の性能改善

第 4 章で述べたように、現在の名寄せプログラムの性能は、充足率、正解率ともかなり満足すべきレベルに達しているが、更に改善が望まれる問題として次の 2 点がある。

- SUBORG サブフィールドへの所在地データ等の混入:この問題については 3-3-1 に述べたとおりで、SUBORG+ORG, ORG+SUBORG マッチの導入でも解決できない表記がある。現在も検討を続けている状況である。
- 大学の第 3 階層以下の組織による表記:この問題については 2-5-2(2)、3-3-2 で述べ、下位機関統計辞書、ユーザー定義統計辞書を用いてある程度の解決を得ていることも示した。しかし、統計的方法で対応しているため、同定した下部組織が間違っていることがある。特に、ある第 3 階層組織が変遷関係にある 2 つの組織の一方に結びつけられているとき、そうでない組織の方が正解である場合があり得る。

(5) 国際的機関リポジトリとの提携

Research Organization Registry (ROR)は、世界の研究機関の識別とそれらのメタデータのオープンな提供を目的とするレジストリーであり、この種の情報源として最も国際的に知られている [16]。California Digital Library, Crossref, DataCite, 及び Digital Science 社が共同で運営している。このレジストリーと連携して、ROR の機関 ID と機関名辞書の NID を相互に収録すれば、機関名辞書の国際展開に有効と考えられるが、業務量の関係からそこまで至っておらず、両者の対応テーブルを一度公開したに留まっている [2]。

以上の課題のうち(3)と(4)は内部での検討を進めるとして、それ以外については、現状の体制、予算の中での努力でできることは限られており、次の 2 つの方向を模索する必要がある。

第一は、関係機関との連携の強化である。上記(1)に関しては、毎年度、32 大学に更新データの確認・修正を依頼して、収録情報の信頼性向上を図っているが、このような関係を更に拡げ、かつ深めていきたい。組織の変遷は頻繁なので、NISTEP のみでこれを追跡することはなかなか難しい。上記(2)については、名寄せプログラムの主要な利用層であるリサーチ・アドミニストレーターの方々の意見と助力を得ていきたいと考えている。(2)の中で述べた日本語機関データの名寄せについては、機関名辞書への日本語別名・揺らぎ名の収録が必要であり、日本語のデータベースを扱っている機関との協力により前進する可能性がある。また、上記(5)については、関係する機関との共同作業とするのが望ましい。

第二は、機関名辞書のためのデータ収集業務(1-4-1 参照)、及び名寄せ結果のチェック業務(第 4 章参照)の大幅な効率化・省力化である。これらについては、業務の委託先も含めて相当な人手をかけている。これらについて部分的に試みているが、AI 的な手法を含めて、今後検討を進めたい。

謝辞

機関名辞書の整備・更新及び名寄せプログラムの開発・改善に携わっておられるリアクトン株式会社の服部正泰社長には、原稿を読んでいただいていくつかの貴重な助言を得た。ここに謹んで謝意を表する。

参考文献

1) “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 事業）”. 文部科学省. https://www.mext.go.jp/a_menu/kagaku/kihon/1348022.htm (参照 2022-08-29)

2) “大学・公的機関における研究開発に関するデータ. ” 科学技術・学術政策研究所. http://www.nistep.go.jp/research/scisip/randd-on-university (参照 2022-08-29)

3) “NISTEP 大学・公的機関名辞書 ver.2022.1. ” 科学技術・学術政策研究所. 2022 年 6 月. http://doi.org/10.15108/data_rsorg001_2022_1 (参照 2022-08-29)

4) “WoSCC-NISTEP 大学・公的機関名辞書対応テーブル ver.2020.1. ” 科学技術・学術政策研究所. 2021 年 1 月. http://doi.org/10.15108/data_rsorg003_2020_1 (参照 2022-08-29)

5) “Scopus-NISTEP 大学・公的機関名辞書対応テーブル ver.2018.1.1 . ” 科学技術・学術政策研究所. 2019 年 12 月. http://doi.org/10.15108/data_rsorg004_2018_1 (参照 2022-08-29)

6) “NISTEP 機関同定プログラム公開版の利用者募集. ” 科学技術・学術政策研究所. 2021 年 11 月. https://www.nistep.go.jp/archives/49078 (参照 2022-08-29)

7) 小野寺夏生. 大学・公的機関における研究開発に関するデータの整備－ミクロデータ分析への貢献－. NISTEP NOTE No.11. 科学技術・学術政策研究所, 2014 年 5 月. http://hdl.handle.net/11035/2926 (参照 2022-08-29)

8) 小野寺夏生, 伊神正貫, 阪彩香. 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 No.15. 科学技術・学術政策研究所, 2015 年 10 月. http://hdl.handle.net/11035/3085 (参照 2022-08-29)

9) 小野寺夏生, 伊神正貫, 富澤宏之. 客観的根拠（エビデンス）に基づく政策のためのデータ・情報基盤（第二回）～NISTEP 大学・公的機関名辞書～. STI Horizon. 2018, vol. 4, no. 3, p. 54-59, http://doi.org/10.15108/stih.00147 (参照 2022-08-29)

10) 小野寺夏生, 中山保夫, 伊神正貫, 富澤宏之. NISTEP の大学・公的機関名辞書と企業名辞書, 及びそれらの識別子. 情報の科学と技術. vol. 71, no. 8, p. 372～375, https://doi.org/10.18919/jkg.71.8_372 (参照 2022-08-29)

11) 中山保夫, 富澤宏之. 客観的根拠（エビデンス）に基づく政策のためのデータ・情報基盤（第一回）～NISTEP 企業名辞書～. STI Horizon. 2018, vol. 4, no. 2, p. 47-53, http://doi.org/10.15108/stih.00134 (参照 2022-08-29)

12) “NISTEP 企業名辞書 ver.2021_1.” 科学技術・学術政策研究所. 2021 年 9 月. https://doi.org/10.15108/data_compdic001_2021_1 (参照 2022-08-29)

13) "the NISTEP Dictionary of Names of Universities and PublicOrganizations ver2022.1." National Institute of Science and Technology Policy. August 2022, http://doi.org/10.15108/data_rsorg001_2022_1_E (参照 2022-08-29)

14) “大学・公的機関名表記ゆれテーブル ver2019.1.” 科学技術・学術政策研究所. 2019 年 7 月. http://doi.org/10.15108/data_rsorg002_2019_1 (参照 2022-08-29)

15) 西川開, 黒木優太郎, 伊神正貫. 科学研究のベンチマーキング 2021. 調査資料-312, 科学技術・学術政策研究所, 2021 年 8 月. https://doi.org/10.15108/rm312(参照 2022-10-18)

16) "Welcome to the Research Organization Registry Community." ROR Community Advisory Group. https://ror.org/ (accessed 2022-09-07)

付録:この報告書で使用する用語について

項番	用語	簡単な説明	初出章節／詳しい説明のある章節
1	代表機関;下部組織	独立した機関を「代表機関」、代表機関に付属する組織を「下部組織」という。	1.1; 1.3
2	機関	代表機関と下部組織を合わせて「機関」という。	1.1
3	NID	機関名辞書における機関の識別キー	1.1
4	正式名(Formal)	機関の正式の名称。日本語正式名と英語正式名がある。	1.2.2; 1.2.3
5	別名(Alias)	正式名以外によく使われる英語の名称や略称。	1.2.3; 3.1.3
6	揺らぎ名(Variant)	名寄せのために機関名辞書に収録される種々の英語名称表記。	1.2.3; 3.1.1
7	非使用名(NotUse)	機関の英語正式名や略称であるが、名寄せには使用しない英語名称。	1.2.3; 3.1.3; 3.4.3; 3.5.2
8	機関の階層	代表機関の階層を"1"、代表機関直下の各組織の階層を"2"、第 2 階層組織の直下の組織を"3"とする。現在の機関名辞書には階層"4"までが存在する。	1.2.4
9	変遷情報	現存しない機関に付与される、移行の区分、移行の年月日、継承機関の情報。	1.2.5; 3.2; 3.4.1; 3.4.2
10	継承機関;継承関係	非現存となった機関の事業等を主に引き継いだ機関を「継承機関」といい、被継承機関－継承機関の関係を「継承関係」という。	1.2.5; 3.4.1
11	大学下部組織種別	大学の下部組織に対する「学部」、「大学院」、「研究所」等の種別。	1.2.7
12	拠点	共同利用・共同研究拠点または WPI の拠点に指定された大学下部組織	1.2.7
13	所管・被所管関係	省庁とその所管する国立研究所や国立研究開発法人等の間、または都道府県庁とその所管する公設機関等の間に、これらの間の同時同定を防ぐために付ける関係。	1.2.7; 3.5.3

項番	用語	簡単な説明	初出章節／詳しい説明のある章節
14	32 大学	機関名辞書に第 2 階層組織を包括的に収録する大学。主に発表論文数の多い大学であるが、下部組織更新情報の提供において協力を得られた大学も含む。	1.3.2
15	ORG; SUBORG; ADDRESS	WoSCC においてそれぞれ代表機関、下部組織、全アドレスデータが記入されたサブフィールド。Scopus については 2.1.4 を参照。	2.1.1; 2.1.4
16	アドレスフィールド	WoSCC 及び Scopus において機関、所在地、郵便番号を示すフィールドをまとめてこのように呼ぶ。	2.1.2
17	略記辞書	名寄せの前処理として単語とその種々の略記を統一するための辞書。	2.2; 2.4(5)
18	下位機関統計辞書;ユーザー定義統計辞書	32 大学の第 2 階層組織に相当する表記が脱落した名寄せ対象データを下部組織同定するための補助ファイル。	2.2; 2.5.2(2); 3.3.2
19	特別措置機関統計辞書	類似名称を持つ大学ペア（3 つ組の場合もある）の間の誤同定を防ぐための補助ファイル。	2.2; 2.5.3(1); 3.5.1
20	大学・短期大学のペア定義テーブル	大学とその大学に属する短期大学部のペアのテーブルで、これらの短期大学部が大学に同定されるのを防ぐために用いる。	2.2; 2.5.2(1); 3.3.3
21	特別ルール	同一または類似の英語名を持つ機関間の誤同定または同時同定を防ぐために個々に定めたルール。	2.2; 2.5.3(2); 3.4.3; 3.5.2; 3.5.3
22	パターンマッチング	上下関係や変遷関係にある機関が同時同定されたときに、最も適切な機関に絞り込むためのルール集。	2.2; 2.5.4(4); 3.2
23	最長マッチ	同定対象データの文字列に最長マッチする機関名辞書中の名称データを持つ機関に同定する方法。	2.3; 2.5
24	混合マッチ	最長マッチで同定できなかった場合に行う同定法で、郵便番号マッチと曖昧マッチの両者でマッチした機関に同定する。	2.3; 2.6.1
項番	用語	簡単な説明	初出章節／詳しい説明のある章節
25	郵便番号マッチ	同定対象データ中に含まれる郵便番号にマッチする番号を持つ機関名辞書中の機関に同定する方法。	2.3; 2.6.1
26	曖昧マッチ	1 文字違いを許容する N-Gram 文字列マッチに適合する機関に同定する方法。	2.3; 2.6.1
27	ベクトルマッチ	最長マッチと混合マッチのいずれでも同定できなかった場合に行う同定方法。同定対象データから変換されたワードベクトルと最も類似する名称データワードベクトルを持つ機関名辞書中の機関に同定する（類似度が所定の閾値を越えて	2.3; 2.6.2

		いる場合に限る)。	
28	前処理	名寄せのマッチング処理を行う前に単語や単語列を正規化する処理。	2.4; 3.1.2; 3.1.4
29	SUBORG+ORG(マッチ)	SUBORG 単語列の後に ORG 単語列を接続したもの。(SUBORG+ORG に対する最長マッチ)	2.5.1; 3.3.1; 3.6.1
30	ORG+SUBORG(マッチ)	ORG 単語列の後に SUBORG 単語列を接続したもの。(ORG+SUUBORG に対する最長マッチ)	2.5.1; 3.3.1; 3.6.1
31	再帰同定	SUBORG+ORG マッチ及び ORG+SUBORG マッチにおいて、一旦同定が行われた後、残った単語列を抽出してもう一度マッチング処理を行うこと。	2.5.1; 3.7
32	同時同定	1 サイクルの同定において複数の機関が同定されること。主に再帰同定により生ずる。	2.5.1; 2.5.4(4); 3.7
33	2 機関比較法	何度も変遷を重ねたためパターンマッチングによる同定機関絞り込みができない場合、2 機関ずつのペア比較により絞り込む方法。	2.5.4(5); 3.2

調査担当

本調査は以下の担当により実施した。

文部科学省科学技術・学術政策研究所

（調査実施、報告書執筆）

小野寺　夏生　　科学技術予測・政策基盤調査研究センター 客員研究官

（調査の総括、報告書のチェック）

伊神　正貫　　科学技術予測・政策基盤調査研究センター センター長

（2022 年 11 月末時点）

(裏白紙)

NISTEP NOTE（政策のための科学）
No. 25

NISTEP における大学・公的機関名辞書の整備と名寄せプログラムの開発
－より精確な研究機関同定（名寄せ）を目指して－

2023 年 1 月

文部科学省 科学技術・学術政策研究所
科学技術予測・政策基盤調査研究センター

〒100-0013 東京都千代田区霞が関 3-2-2 中央合同庁舎第 7 号館 東館 16 階
TEL: 03-6733-4910 FAX: 03-3503-3996

Construction of the Dictionary of Names of Universities and Public Organizations and development of the Program for Organization Name Disambiguation in NISTEP: Toward more precise and accurate research organization name disambiguation

January 2023

Center for S&T Foresight and Indicators
National Institute of Science and Technology Policy (NISTEP)
Ministry of Education, Culture, Sports, Science, and Technology (MEXT), Japan

https://doi.org/10.15108/nn025

https://www.nistep.go.jp