STI Horizon 2023 Vol.9 No.3
https://doi.org/10.15108/stih.00339

レポート

機関名辞書及び名寄せプログラムとその補助ツール －それらの概要と利用の留意点－

科学技術予測・政策基盤調査研究センター
客員研究官 小野寺 夏生、客員研究官 吉井 隆明

概　要

NISTEP 大学・公的機関名辞書と、この辞書を用いた名寄せプログラムの構成、機能、利用上留意する点について概説する。これらは、機関レベルあるいは組織レベルでの研究データ分析を精確に行う上での不可欠な手段として利用されることを目指している。本稿は、2023 年 1 月に公表された NISTEP NOTE 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－」に基づくが、その後の辞書の改善や名寄せのための補助ツールの開発についても触れる。

キーワード：大学・公的機関名辞書，名寄せ，研究機関，研究開発動向，データ分析

はじめに

多種多様な情報源を用いて機関レベル、組織レベルで研究開発データを整理・分析しようとすると、機関名の揺らぎ、下部組織情報の不足などの問題点に直面する。NISTEP 大学・公的機関名辞書（以降「機関名辞書」、あるいは単に「辞書」という）と、この辞書を用いた名寄せプログラムは、これらの問題に対処して、精確な機関同定を行う手段の提供を目指すものである。

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[1,2]、リサーチ・アドミニストレーター（RA）や研究開発分析を行う研究者等に活用されている。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機関同定に NISTEP 内部で利用し、その結果を公開しているが、2021 年度からプログラム自身の公開を開始し、現在 30 名以上の利用者がある。

本稿は、機関名辞書と名寄せプログラムについて、特にその利用者に留意してほしい点を重点に述べたものである。より詳しくは、これに関する NISTEP NOTE[3] を参照されたい。

1. 機関名辞書の概要

1.1　収録対象とする機関

収録対象は、研究開発を行っている国内機関である。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の特徴として第一に挙げられるのは、独立した機関（「代表機関」という）のほか、その主な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。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。

下部組織中特に網羅的に収録するのは以下の機関である。

① 大学の下部組織のうち、附属病院、国立大学の附置研究所、及び「拠点」（共同利用・共同研究拠点又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）は、階層のいかんに関わらず収録する。

② 44 大学[注1] については、事務的組織以外のすべての第 2 階層組織（必要に応じて第 3 階層以下の組

織も）を収録する。

③　4 つの大学共同利用機関に属する研究所、大規模な国立研究開発法人の主要な研究所、国の機関及び独立行政法人（認可法人を含む）に属する病院及び大学校病院は必ず収録する。

特徴の第二は、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。

1.2　収録する情報

機関同定の精確性及び組織の改組や再編等も捉えることができるように、①機関 ID（個々の機関に一元的に割り当てた ID）、②機関のセクターと病院フラグ、③機関の日本語名称、④機関の英語名称、⑤機関の階層関係、⑥機関の変遷情報、⑦機関の郵便番号、⑧大学の下部組織種別、⑨外部の機関識別情報、を収録している。

②の「機関のセクター」では、収録する機関を、国立大学、私立大学、非営利団体等 17 のセクターに分類している。④の「機関の英語名称」では、精確な名寄せを図るために英語名称が正式名なのか、別名なのか、揺らぎ名なのか、非使用名なのかの 4 種に区分して収録している。⑥の「機関の変遷情報」では、機関が非現存となった年月日、非現存となった理由（「統合」、「廃止」、又は「変更」）を記入し、継承機関があればその機関名を収録している。⑨の「外部の機関識別情報」には、NISTEP 企業名辞書[4, 5] の企業 ID、大学・高等専門学校コード、科研費機関番号を収録している。

1.3　機関名辞書の公開

機関名辞書は、2012 年 12 月に初めて公開し、それ以降年 1、2 回更新している。2023 年 7 月の公開版では、総数 21,205 機関、そのうち 16,323 が代表機関、4,882 が下部組織である。また、全機関中現存するのは 14,174 機関（代表機関 10,622、下部組織 3,552）である。また、2021 年 1 月に初めての英語版を公開し、その後日本語版と合わせて更新している[6]。

2. NISTEP で実施している名寄せの方法

機関同定の対象となるのは、国内の研究機関の英語表記である。従来の主な対象である WoSCC と Scopus のデータの名寄せに傾注してプログラム開発を行ってきた。しかし、国内機関に使われる英語の別名、揺らぎ名はかなり共通なので、他のデータ源にも一般に適用可能である。

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。以下では、レコード内の機関データは、代表機関名を示すフィールド（ORG）、下部組織名を示すフィールド（SUBORG）、機関とそのアドレスの全情報を示すフィールド（ADDRESS）等に分割されている前提で説明を行う。

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、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、又はキャメルケース（複合語やフレーズを表記する際、各単語や要素語の先頭の文字を大文字で表記する手法）で表記された語（"RadioIsotope" 等）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略記辞書を用いた語の正規化（"Science", "Sciences", "Scientific" → "Sci" 等）、⑥一部語尾の処理、を行う。

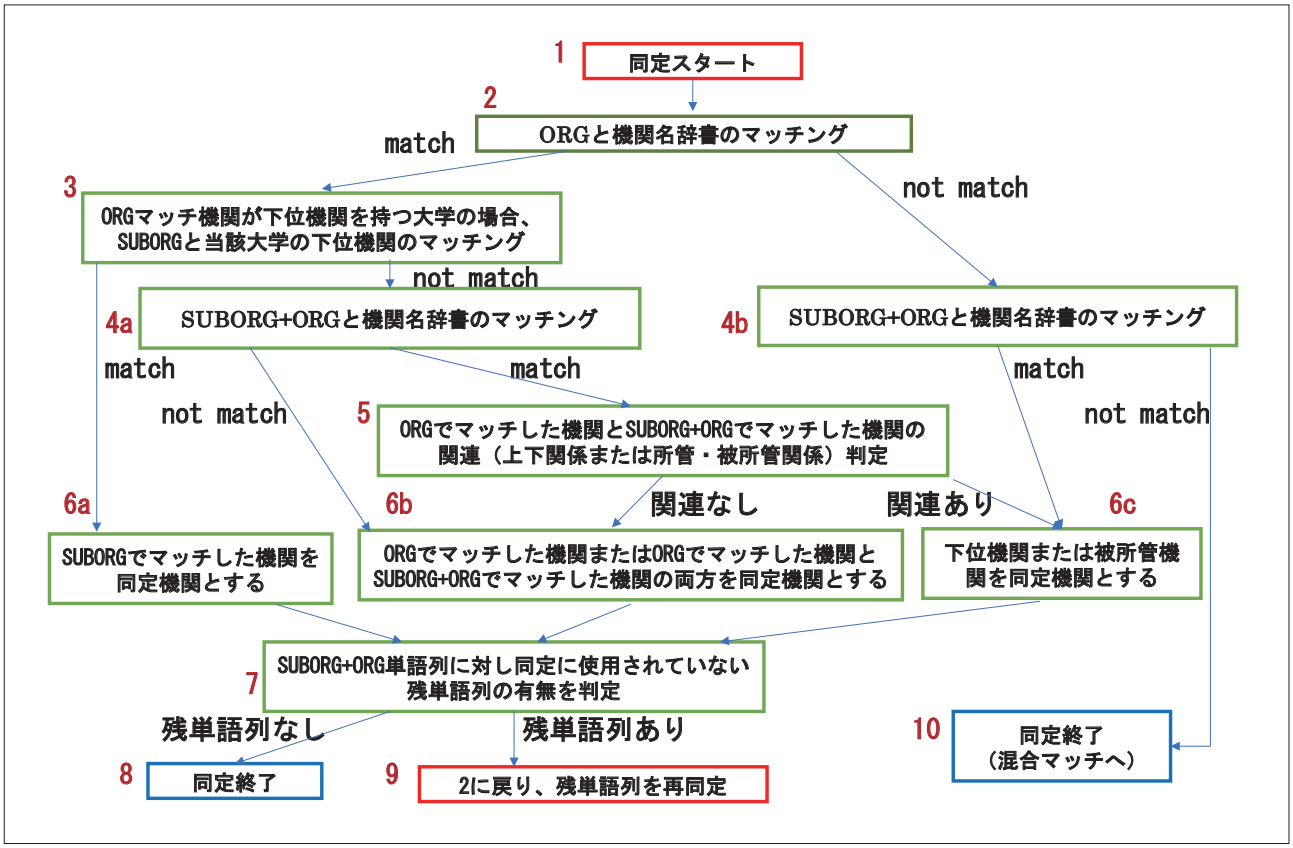
2.2　最長マッチ

前処理を行った単語列に対し、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。図表 1 がその流れ図である。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 9 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）。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。しかし、"Div Mammalian Dev, Natl Inst Genetics, Dept Genetics, SOKENDAI " のように、1 つの所属機関情報の中に複数の機関（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）が記入されるこ

注 1　この 44 大学のほとんどは WoSCC データベースの収録論文数が多い大学であるが、下部組織情報提供に協力を得ている少数の大学を含む。具体的には参考文献 1）や 2）の利用マニュアルを参照。

図表１　最長マッチの流れ



とがある。再帰同定は、主にこのような場合に対応するためである。

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（なるべく長い文字列長でレーベンシュタイン距離 (2 つの文字列がどの程度異なっているかを示す距離の一種）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とする。

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。その主要な方法はパターンマッチングで、同定候補に挙がった機関が上下関係にある場合はより下位の機関を優先する等のルールを定める。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、いずれも同定とする。

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に対しては、①セクター判定、②病院であるか否かの判定を行い、いずれの判定もされなければ③同定不能とする。

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

(1) 44 大学の下部組織同定に用いるサポート辞書

1.1 に述べた 44 大学については第 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" を省略した表記がしばしば見られる。このような場合、機関名辞書の下部組織名とはマッチしないので、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「下位機関統計辞書」等のサポート辞書を用いる。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。

(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複雑多様で、ORG と SUBORG に分離していることが多い。

② SUBORG サブフィールドには、下部組織の名称だけでなく、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、図表 1 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）。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、検討を続けているところである。

3.2 機関（組織）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。これらに対しては以下の対策を講じている。

(1) 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 参照）。英語名変更後の論文で旧名がよく用いられるときには、これらの旧名を別名又は揺らぎ名として登録する。

(2) 変遷前後で英語名が変わらない場合は、同定対象のデータの発行年と、機関名辞書に記録された移行年を比較し、発行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 英語名が同一又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。このようなときに正しい機関に同定するため、以下の対策を講じている。

(1) 機関名辞書での対応

類似の名称を持つ機関を機関名辞書に登録する。例えば、旧・国立公衆衛生院の英語名 The Institute of Public Health は、多くの地方自治体の研究所の名称と似ている（Aichi Prefectural Institute of Public Health 等）。これらの名称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。

(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) 特別ルールの設定

一方の機関の ADDRESS 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、一例のみを挙げる。

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の英語名は Institute for Environmental Sciences であるが、同じ単語列を含む機関が、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。そこで、ADDRESS フィールドに "Aomori", "Rokkasho" 等のいずれかが存在する場合に限りこの機関に同定するルールとする。

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) 機関の略称、(d) スペル方式の違い、等がある。(b) と (d) に対しては、主に 2.1 で述べた前処理において対処する。(a) と (c) に対しては、よく使われる略称や揺らぎ名を別名や揺らぎ名にする等、機関名辞書で対応する。

4. 同定漏れと誤同定の検出と対処

名寄せプログラムにより同定されなかった結果あるいは同定され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が必要である。このため、一定以上の出現頻度があった ADDRESS 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、対策をとっている。

4.1 機関同定できなかったデータの調査

近年の WoSCC や Scopus の機関同定では、処理されたレコードに対する機関同定されたレコードの割合（充足率）は 93～95% である。未同定データに対してその理由を検討し、辞書への新たな機関の登録、既収録機関への揺らぎ名の追加、名寄せアルゴリズムの修正や特別ルールの設定を行う。しかし、これ

らにより充足率を上昇させることは正解率の低下を伴いがちであるので、誤同定の発生を招かない範囲で行い、特に処置は行わないことも多い。

4.2　主な誤同定の内容とそれへの対処

目視チェックの結果同定が誤っているデータに対しては、その理由を検討し、機関名辞書の修正（機関の新規登録、揺らぎ名の追加や削除等）又はプログラムの修正（特別ルール設定等）により解決を図る。2020 年度に行った WoSCC データの機関同定をチェックした結果ではエラー率は 3.3% であったが、その大部分は、下部組織とその代表機関の間、又は変遷前後の機関間の間違いであり、重大である代表機関に関する誤りは 0.04% であった。この名寄せの精確度は完全ではないがかなり高いと言える。以降の名寄せでは 99.5% 以上の正解率が達成されると予想しているが、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から、エラーの検出と対策検討は今後も必要である。

5. 公開プログラム利用者のための利用ツール

より広い関係者への名寄せプログラムの利用拡大を目指して、2020 年度及び 2021 年度に希望者を募り利用者の声を拾った。

2020 年度は、公開に向けた準備の試用実験の位置づけで、試用者を募って約 4 か月試用していただき評価を得た（参加者 41 名、評価回答者 14 名、試行版継続利用希望 12 名）。

2021 年度は、NISTEP 機関同定プログラム公開版としての利用者を募り、約 6 か月使用していただき利用者の声をアンケートで拾った（参加者 29 名、アンケート回収 14 名）。

また、2022 年度には、RA 協議会第 8 回年次大会ポスター発表で、2020 年度及び 2021 年度の利用者の声を取りまとめて公表し、ポスターセッション参加者の声をアンケートで拾った（回答者 28 名）。

以上の利用者の声を踏まえて、優先的に開発したツールは以下の 2 つで、入力ファイル成型ツールは 2023 年 2 月に公開し、辞書更新 Web ツールは 2023 年度の公開を予定している。

〇入力ファイル成型ツール：

このツールは、WoSCC あるいは Scopus の Web サイトよりダウンロードした論文データファイルを、名寄せプログラムの入力ファイル形式に変換するプログラムである。これにより利用者が希望する論文ファイルで機関同定が可能となる。

〇辞書更新 Web ツール：

このツールは、機関名辞書の利用者が、Web 上で辞書更新の操作ができるプログラムである。このツールによって、利用者は機関名辞書に独自の機関名の追加登録や収録済み機関への情報（英語別名等）追加等（辞書更新機能）ができるようになる。また、公開中若しくは公開済みの辞書ファイルの中から年版を選択してダウンロード可能となることから、利用者は論文発表年に応じた辞書ファイルを利用して機関同定が可能となる。

おわりに－未解決の課題

最後に、代表的な今後の課題を述べる。

(1) 2021 年度まで、下部組織を包括的に収録している大学は 32 大学に過ぎなかったが、2022 年度に 44 大学に拡張した。本事業の持続可能性も踏まえつつ、今後も可能な範囲で拡大を図る。

(2) 名寄せプログラムの利用者から種々の要望があり、これらに対して順次対応を考えている。5. に述べたように、要望を踏まえていくつかのツールを開発・導入した。

(3) 3.1 で述べたように、下部組織の名寄せは、大学、大学以外とも問題が残っている。これらに対しては適時対応を行っているが、連合大学院や各種の機構など、本事業開始時点では想定していなかった事例も出てきているので、これらも含めてより適切な方策を考えたい。

これらの未解決の課題に対応するには、機関名辞書や名寄せプログラムの利用者の協力やフィードバックも必要であると考えている。利用希望の際は、NISTEP 担当宛（noip-registration[at]nistep.go.jp ([at] を "@" に変更してください)）への連絡をお願いしたい。

参考文献・資料

1)　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,

https://doi.org/10.15108/stih.00147（参照 2023-05-23）
2) “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.
https://doi.org/10.15108/data_rsorg001_2022_1（参照 2023-05-23）
3) 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 No.25. 科学技術・学術政策研究所，2023 年 1 月.
https://doi.org/10.15108/nn025（参照 2023-05-23）
4) 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,
https://doi.org/10.15108/stih.00134（参照 2023-05-23）
5) “NISTEP 企業名辞書 ver.2022_1.”科学技術・学術政策研究所．2022 年 11 月.
https://doi.org/10.15108/data_compdic001_2022_1（参照 2023-05-23）
6) “the NISTEP Dictionary of Names of Universities and Public Organizations ver2022.1.” National Institute of Science and Technology Policy. August 2022,
https://doi.org/10.15108/data_rsorg001_2022_1_E（参照 2023-05-23）