NISTEP NOTE（政策のための科学） No.11

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－

2014 年 5 月

文部科学省　科学技術・学術政策研究所

科学技術・学術基盤調査研究室

客員研究官　小野寺　夏生

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に取りまとめた資料です。

NISTEP NOTE (Science of Science Technology and Innovation Policy) No. 11

Development of Data Infrastructure on R&D Activities in Universities and Public Organizations
– NISTEP's Contribution to Micro-Data Analysis –

Natsuo ONODERA

May, 2014
Research Unit for Science and Technology Analysis and Indicators,
National Institute of Science and Technology Policy (NISTEP)
Ministry of Education, Culture, Sports, Science and Technology (MEXT)
Japan

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－

小野寺夏生[1]
[1]文部科学省　科学技術・学術政策研究所　科学技術・学術基盤調査研究室

要旨

研究開発の動向を包括的に把握するためには、研究開発のインプット及びアウトプットに関する多様なデータを収集・加工・組織化する必要がある。論文データベースから得られる研究アウトプットデータの機関やその下部組織レベルでの分析(ミクロデータ分析)には、様々に表記される著者所属機関や下部組織の名称を正しく同定する作業が伴う。

NISTEPでは、研究者や政策担当者によるミクロデータ分析を支援するために、「大学・公的機関に関するデータ整備」を2011年度より実施している。本プロジェクトを通じて、機関同定の核となる機関名辞書や、論文データベースにおける日本の大学および公的機関の表記ゆれリストなどを整備し、公開している。本報告書では、このプロジェクトの概要を述べるとともに、論文データベースにおける所属機関名の表記ゆれの実態について示す。それを踏まえ、機関名記述方法の統一の必要性について述べる。

Development of Data Infrastructure on R&D Activities in Universities and Public Organizations
- NISTEP's Contribution to Micro-Data Analysis -

Natsuo Onodera[1]
[1] Research Unit for Science and Technology Analysis and Indicators, National Institute of Science and Technology Policy (NISTEP), MEXT

ABSTRACT

For comprehensively understanding the status and trends of R&D activities in a country or a region, various data on research input and output should be collected, processed, and organized. Specifically, data analysis of research output data obtained from bibliographic databases at the organizational or departmental level (micro-data analysis) is necessarily accompanied with accurate identification of author-affiliated organizations and departments which generally have numerous name variations.

In order to help micro-data analysis conducted by researchers and policy-makers, NISTEP has carried out a project "Development of data infrastructure on R&D activities in universities and public organizations" since FY2011. Through this project, it prepares and publishes an organization name dictionary playing a central role in identification and some lists of name variations in databases for universities and public organizations in Japan. This report outlines the project, with some results of analysis on name variations of author-affiliated organizations. Finally, it discusses importance of standardization of organization name description.

（裏白紙）

目次

＜概要＞

（裏空白）

概要

1. 本報告書の概要

研究開発の動向を包括的に把握するためには、研究開発のインプット及びアウトプットに関する多様なデータを収集・加工・組織化する必要がある。論文データベースから得られる研究アウトプットデータの機関やその下部組織レベルでの分析(ミクロデータ分析)には、様々に表記される著者所属機関や下部組織の名称から、それが示している機関または組織を同定する作業(これを機関の名寄せという)が伴う。

本報告書では、まず論文データベースの一般的な構成を示し、これをデータ源としてミクロデータ分析を行う際の問題点、注意点について考察する。次に、NISTEP において 2011 年度から実施している「大学・公的機関に関するデータ整備」事業の概要を述べ、そこで整備・公開しているデータが、ミクロデータ分析活動をどのように支援するかを示す。その後、本データ整備事業の中で行った論文データベースにおける機関名表記のゆれの分析結果を報告する。

2. 大学などの機関ごとの状況を把握する分析の難しさ

大学や公的機関の研究アウトプットに関するデータについては、論文データベースが主要なデータ源となるが、これによって研究開発の実態や動向を正確に分析するためには、十分なデータの整理、クリーニングが必要である。特に、ミクロデータ分析で重要な所属機関データについては、同じ機関の名称が論文により様々に表記されること、いわゆる「表記のゆれ」の問題があり、このため機関の名寄せが必須になる。

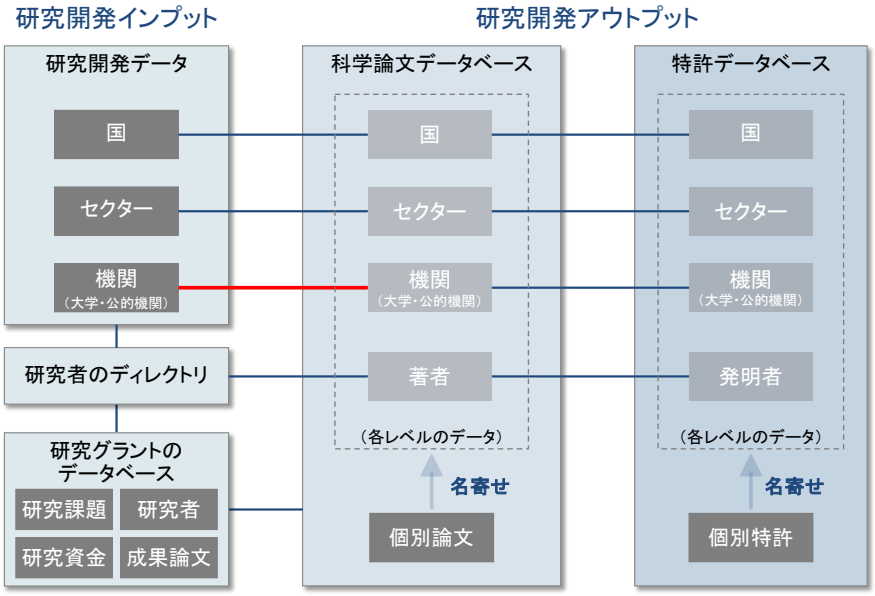
この他、下部組織名の表記の多様性、機関や組織の変遷、同一著者の複数機関所属等も、機関名寄せを行うとき厄介な問題となる。

3. 大学・公的機関に関するデータ整備－NISTEP における取組み

NISTEP では、文部科学省の「科学技術イノベーションにおける“政策のための科学”推進事業」の一環として、平成 23(2011)年度から「データ・情報基盤の構築」を実施しており、その中のプロジェクトの一つが「大学・公的機関に関するデータ整備」である。これは、我が国における研究開発（特に政府予算で実施されているもの）の実態の把握・分析及びそのパフォーマンス評価を、国、セクター、個別機関などの各レベルで行うための基礎として、大学・公的機関の科学技術生産に関するデータの整備を行うことを目的としている。

概要図表 1 は、その構想をモデル的に示したものである。この図表のうち、国レベルのデータ集計やインプットとアウトプットのデータ接続（マクロデータ分析）は比較的容易であるが、セクターレベルの分析（メゾデータ分析）や機関レベル及び研究者レベルの分析（ミクロデータ分析）には困難が伴う。その主な理由は、論文データベースや特許データベースにおいて機関や研究者の名寄せ、機関のセクター同定が必要なことによる。

概要図表 1　大学・公的機関における研究開発に関するデータ整備の概念モデル



「大学・公的機関に関するデータ整備」では、特にミクロデータ分析に必要な基盤データやツールの開発に注力しており、ここで整備したデータは、関係者に利用していただくため、次に示す NISTEP の「データ・情報基盤」の Web サイトにおいて公開を進めている。

http://www.nistep.go.jp/research/scisip/randd-on-university

現在、次の(1)と(2)に関するデータが公開されている。これらは、ミクロデータ分析、その他日本の研究機関に関する分析に際し正確で高精度な機関同定を行うための活用が期待される。

(1) NISTEP 大学・公的機関名辞書の整備

NISTEP 大学・公的機関名辞書(以下単に「機関名辞書」という)は、インプットデータ、アウトプットデータを機関レベル及びセクターレベルで分析するための基本情報を含む。収録対象は研究開発を行っている国内の機関で、大学、公的機関を重点とするが、地方公共団体の機関、企業、非営利法人等もできるだけ含めており、全部で 10,000 機関以上に達している。それぞれの機関には NISTEP 独自の識別 ID を与え、以下の情報を収録する。

① 機関の名称:和英の正式名称の他、英語名については、通称、略称もできるだけ収録。
② セクター:**概要図表 2** に示す 16 のセクターに各機関を分類。
③ 機関の下部組織:主要な大学、大学共同利用機関、独立行政法人に属する下部組織を収録。
④ 機関の変遷情報:統廃合、改組、名称変更等の情報をできるだけ収録。

現在、2012 年度末時点での機関名辞書を前記 Web サイトで公開しており(NISTEP 大学・公的機関名辞書(Ver.2012.1))、2014 年度には、データ拡充を行った改訂版を公開予定である。

概要図表 2 機関名辞書で使用するセクターとセクターごとの収録機関数

セクター	収録機関数	セクター	収録機関数
国立大学	101	私立高専	3
国立短大	26	大学共同利用機関	5
国立高専	59	国の機関	135
公立大学	94	特殊法人・独立行政法人	133
公立短大	62	地方公共団体の機関	696
公立高専	6	会社	4,421
私立大学	601	非営利団体	3,586
私立短大	515	その他の機関	6
		計	10,449

(2) 論文データベースにおける機関名寄せ:機関名辞書とのミクロ接続

Scopus と Web of Science Core Collection (以下 Web of Science または WoS)から、1996～2011 年の期間に発表された日本の論文(日本の機関に属する著者を少なくとも一人含む論文)を抽出し、そこに含まれる機関を名寄せして、機関名辞書の登録機関と対応づけた。Scopus では延べ 329 万件の機関データのうち 91.9%、WoS では延べ 278 万件の機関データのうち 93.6%が機関同定できた(2012 年度末時点)。これらのサンプリング調査により、同定の精度は 98%以上であることを確認した。現在は、これらの結果の評価に基づき機関名辞書と名寄せアルゴリズムの改善を行い、同定率と同定精度の向上を目指している。

2012 年度末時点でのデータ整備に基づき、次のデータを、前述の「データ・情報基盤」サイトから公開している。

① 大学・公的機関名英語表記ゆれテーブル(Ver.2013.1)
② Scopus-NISTEP 大学・公的機関名辞書対応テーブル(Ver.2013.1)

(3) その他のデータ整備活動

この他、以下のデータ整備活動を行っている(将来のデータ公開については検討中)。

(a) 研究インプットデータベースと機関名辞書とのミクロ接続

2002～2011 年度の科学技術研究調査対象名簿の機関名と機関名辞書を対応づけるためのプログラムを開発した。

(b) 論文生産統計のためのテーブル設計

(2)で述べた結果に基づいて、Scopus と WoS のデータに対する種々の論文生産統計が可能なようにテーブル設計を行った。機関別・セクター別に、年別・分野別及びこの両者を組み合わせた集計が、整数カウントと分数カウントにより可能である。

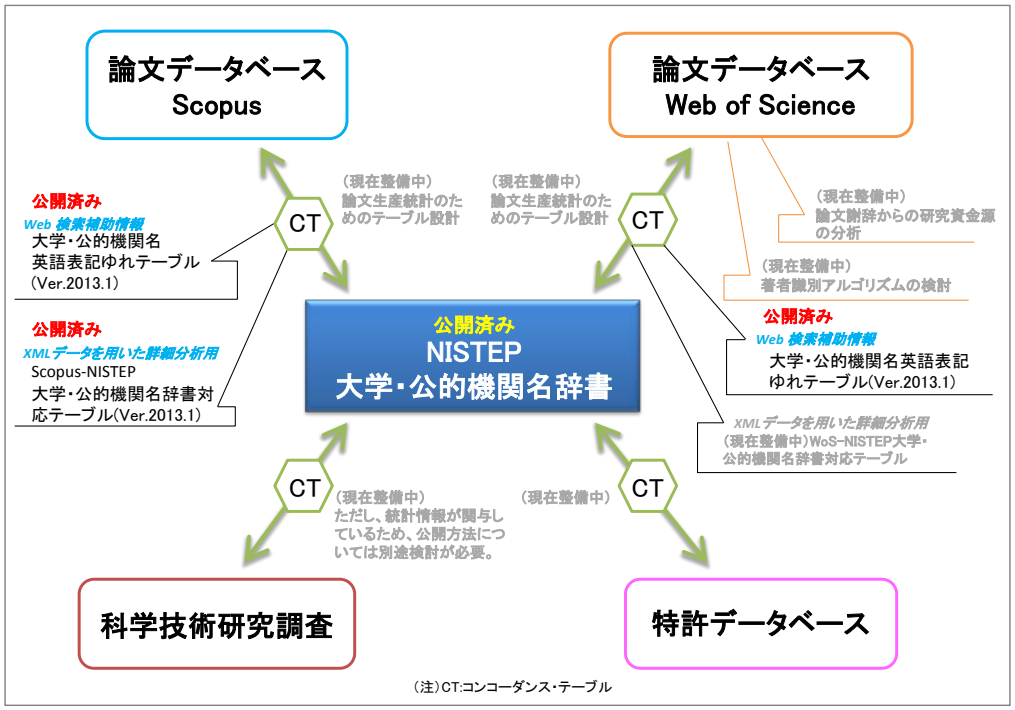
(c) その他

特許データベースと機関名辞書のミクロ接続、WoS の論文謝辞データからの研究資金源の分析、WoS を対象として著者識別アルゴリズムの検討も行っている。

下記の**概要図表 3** は、機関名辞書を中心とした各種研究開発関連データの接続イメージである。機関名辞書を中心にして、多種多様な研究開発関連データが接続され、その結果、多様な変数の組合せによるデータ分析が可能となる。これは、より多くの仮説の検証の機会を提供することを意味する。

機関名辞書は、我が国の大学や公的機関を網羅的に収録していること、Web 上で公開して誰でも自由に利用できることが特徴である。利用をオープンにすることによって、様々な研究者が研究を行う際に、大学・公的機関に関する情報の典拠としての役割を果たすことを一つの目標としている。

概要図表 3　機関名辞書を中心とした各種研究開発関連データの接続



4. 機関名表記ゆれの状況

上記の大学・公的機関に関するデータ整備を進める中で、機関名表記ゆれに対処するための名寄せ作業に多くの時間を費やすこととなった。ここでは、Scopus における機関名表記ゆれの分析を行った結果の一部を述べる。

(1) 表記ゆれの程度が大きい機関の例

東京農工大学と東京薬科大学の例をそれぞれ**概要図表 4、5** に示す。いずれも、先頭に正式の英語名表記を、以下出現頻度 10 以上の表記を頻度の多い順に示している。

概要図表 4 東京農工大学の表記ゆれ

機関名	機関ID	セクター	Scopusにおける表記ゆれ	英語正式名	出現度数
東京農工大学	NID201200980805842	国立大学	Tokyo University of Agriculture and Technology	○	6166
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agric. and Technology		2993
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agric. and Technol.		227
東京農工大学	NID201200980805842	国立大学	Tokyo Noko University		121
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agriculture/Technol.		112
東京農工大学	NID201200980805842	国立大学	Tokyo University of Agriculture and Technology (TUAT)		109
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agric./Technology		79
東京農工大学	NID201200980805842	国立大学	Tokyo University of A and T		67
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. Agric. T.		53
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agriculture and Technology		40
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agri. and Tech.		33
東京農工大学	NID201200980805842	国立大学	Tokyo University of Agric./Technol.		33
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agriculture Technol.		25
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agriculture/Tech.		17
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agric. and T.		15
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agr. and Tech.		14
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agriculture and Tech.		14
東京農工大学	NID201200980805842	国立大学	Tokyo University of Agri. and Tech.		14
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agri. and Technology		13
東京農工大学	NID201200980805842	国立大学	University of Agriculture and Technology		13
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. Agric. and Technology		12
東京農工大学	NID201200980805842	国立大学	Tokyo University of Agriculture and Technology (TAT)		11
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. Agriculture/Technology		10
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of A and T		10

概要図表 5　東京薬科大学の表記ゆれ

機関名	機関ID	セクター	Scopusにおける表記ゆれ	英語正式名	出現度数
東京薬科大学	NID201200689092004	私立大学	Tokyo University of Pharmacy and Life Science	○	1030
東京薬科大学	NID201200689092004	私立大学	Tokyo Univ. of Pharm. and Life Sci.		1106
東京薬科大学	NID201200689092004	私立大学	Tokyo University of Pharmacy and Life Sciences		572
東京薬科大学	NID201200689092004	私立大学	Tokyo Univ. of Pharm./Life Science		248
東京薬科大学	NID201200689092004	私立大学	Tokyo Univ. of Pharmacy/Life Science		90
東京薬科大学	NID201200689092004	私立大学	Tokyo College of Pharmacy		62
東京薬科大学	NID201200689092004	私立大学	Tokyo Univ. Pharm. Life S.		40
東京薬科大学	NID201200689092004	私立大学	Tokyo Univ. of Pharm. and Life S.		33
東京薬科大学	NID201200689092004	私立大学	Tokyo Univ. of Pharmacy/Life Sci.		15
東京薬科大学	NID201200689092004	私立大学	Tokyo Univ. of Pharm./Life Sci.		12

(2) 表記ゆれのタイプ分け

大学における機関名の表記ゆれは次の 6 つのタイプに類別化される。

概要図表 6　機関名の表記ゆれのパターン

表記ゆれのパターン		補足説明
①	正式の名称とは単語や語順が異なる表記	東京農工大学をTokyo Noko University of Technologyと表記するような例はこれに当たる。
②	単語の略記	UniversityをUniv.、Science (Scientific)をSciとするような表記である。
③	機関の略称	東京工業大学をTITとするような表記である。AIST(産業技術総合研究所)、JST(科学技術振興機構)など公的機関に特に多い。
④	冠詞、前置詞、接続詞の省略や書き換え	Universityで始まる大学名(University of Tokyoなど)の先頭に"The"を付ける表記と付けない表記が存在する。名称中の"of"、"for"、"and"等の前置詞や接続詞が省略されて表記されたり、間違って表記されたり(forをofと誤記するなど)することも多い。
⑤	機関の旧名の表記	英語名称を変更した機関において、旧名が使われることがある。
⑥	スペルの違い	単数形と複数形のゆれ(ScienceとSciencesなど)、ローマ字表記のヘボン式と訓令式のゆれ等がある。

(3) 誤同定が起こりやすい表記

機関名に表記ゆれがあると、機関別の集計や分析を行うとき二通りの問題が起こる。一つは、ある機関に該当するデータを網羅的に得られないことである。機関名辞書や表記ゆれテーブルの公開はこの面での支援になると考えられる。もう一つの問題はより重大で、ある機関名を、それと名称が似ている機関に誤同定してしまうことである。誤同定を起こしやすい表記を、いくつかのパターンに分けて示す。

(a) 同一の英語機関名

大学や公的機関では、全く無関係の機関が同一の英語名を持つことはほとんどないが、統合や改組を行った機関が、日本語機関名は変更したのに英語名はそのままという例は珍しくない。たとえば、東京都立大学と首都大学東京(どちらも Tokyo Metropolitan University)、宇宙科学研究所(国立研究所から独立行政法人宇宙航空研究開発機構の下部組織に移行したがどちらも Institute of Space and Astronautical Science)などである。これらの場合、論文に表記された機関が旧機関か新機関か厳密に判別するのは困難である。

(b) 下部組織が結合した表記

機関データに機関名とその下部組織名が合体表記される場合がある。このような表記が、別の機関名と似通っていると誤同定が起こりやすい。特に、前置詞等が省略されていると判別が困難である。東京理科大学では旧名の Science University of Tokyo もよく使われるので、Fac Sci Univ Tokyo を、東京大学理学部か東京理科大学か判別するのは難しい。

(c) ありふれた単語のみから成る機関名

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、誤同定が起こりやすい。たとえば、分子科学研究所(Institute for Molecular Science)と類似した名称の付属研究所を持つ機関は多数存在する。

5. 大学・公的機関に関するデータ整備から得られた示唆

以上を踏まえ、論文執筆の際の所属機関表記について、本データ整備から得られた示唆をまとめる。

論文を発表するときには、著者の所属機関・組織を正確に表記することが求められる。不統一、不正確な表記は、機関や組織の業績評価に不利益をもたらすことになりかねない。

たしかに論文データベース提供機関が、個々の研究機関の名寄せについても積極的に取り組んでいる。しかしながら、大学もしくは部局ごとに今一度英語表記の統一化を図ることで、論文発表に関する意識の向上がなされるのではないだろうか。また、タイムズ社の大学ランキングにおいては論文数のような定量的指標のみではなく、世界の研究者の間での存在感(visibility)に関する定性的な調査結果も含まれている。このような visibility の向上のためにもやはり大学名や部局名を統一化させることが重要ではないだろうか。従って、個々の論文発表者が注意すると同時に、機関全体で統一的表記を定め、構成員にそれを周知徹底することが望ましい。

大学の場合、大学院生や研究員への教育も必要である。現在、多くの大学や研究所では、その構成員の発表論文を機関リポジトリから公開しているので、このような周知・教育には、機関リポジトリの運営に当たっている図書館が関与するのが効率的であるかもしれない。

また、当該の機関の努力だけでなく、論文が発表される学術雑誌においても、正しい統一的表記が受け入れられるように投稿規定を定めることが必要である。

以下に、所属機関表記に当たって特に注意してほしい点をまとめておく。

[1] 機関や組織の正しい名称を正確に表記する。機関名と組織名は明確に分離する(“Faculty Y X University” ではなく、“Faculty of Y, X University”のように)。

[2] 大学の教員が学内の複数の組織に属していることが多いが、論文発表の際はどの組織を記載するか、大学全体で見解を統一することが望ましい。

[3] 著者が 2 つ以上の機関を兼務している場合、研究に外部資金を得ている場合、ある機関から別の機関に派遣されている場合など、著者所属に複数の機関を記載しなければならないことがある。このような場合はそれぞれの機関を分離して記載する。たとえば、X 大学に所属する著者が JST の CREST によって研究を行った場合、“X University, JST CREST”ではなく、“X University”と“CREST, Japan Science and Technology Agency”の 2 つの所属を記載する。

[4] いくつかの大学共同利用機関や独立行政法人では、それらの機構の下にかなり独立性の高い多くの研究所や施設が存在する。このような機構では、機構名と研究所名を併記するか、研究所名のみを記載するかを機構全体で定めた上で、その記法を統一することが望ましい。

＜本編＞

（裏空白）

1　はじめに

研究開発の動向を計量的に分析することは、その実態を客観的に把握するためにも、今後の計画を定めるための基礎データを得るためにも、極めて重要である。それぞれの大学や研究機関においては、自機関のデータを他機関と比較することにより、自機関の特徴や弱点を把握し、今後の研究開発方針や計画の参考とするのに有用であろう。また、科学技術動向に関心を持つ研究者や科学技術政策の担当者にとっては、国内の研究開発構造の把握、海外諸国の主要機関との比較等により、今後の研究開発動向の予測、科学技術政策の立案等に役立つ知見が得られるであろう。

研究開発の状況、動向を包括的に把握し、理解するためには、次のようなデータが必要である。

図表 1　研究開発活動に関するデータの種類

研究開発活動に関するデータ	具体例
インプットに関するデータ	研究者数、研究開発費など
アウトプットに関するデータ	研究成果物（論文、特許等）の生産量、それらの影響度など
その他のデータ	研究者間の交流、研究開発と社会との関係等に関するデータ

これらのデータを組み合わせることによって新たなデータを生み出すこともできる（たとえば、インプットとアウトプットの関係を示すデータ）。

また、これらのデータは、次のような次元から集計・分析できる。

図表 2　集計・分析に用いる次元の整理

集計・分析に用いる次元		具体例
(a)	研究者の所属の次元	個々の研究者、研究機関あるいは研究組織（ミクロデータ分析）
		国、地域、全世界（マクロデータ分析）
(b)	主題分野あるいはトピックの次元	
(c)	時間の次元	経時的変化や年代的特徴

次元(a)に関して、研究者レベル、研究機関(組織)レベルでのデータ分析をミクロデータ分析、国や地域や全世界のレベルでの分析をマクロデータ分析と呼ぶこととする。本報告書では、科学技術・学術政策研究所(NISTEP)で進めているミクロデータ分析のためのデータ整備事業について述べる。この事業では、論文データベースをデータ源とした研究アウトプットデータの機関・組織レベルでの分析に最も力を入れているので、本報告書の内容もそれについての記述が中心である。

第 2 章では、論文データベースの一般的なデータ構成を示した後、これをデータ源としてミクロデータ分析を行う際の問題点、注意点について考察する。第 3 章では、NISTEP において 2011 年度から実施している「大学・公的機関に関するデータ整備」事業の概要を述べ、そこで整備・公開しているデータが、ミクロデータ分析活動をどのように支援するかを示す。第 4 章では、データ整備の中で最も注力している機関名寄せにおける経験から、データベースにおける機関名表記のゆれの分析の一端を報告する。終章の第5章では、所属機関記述についての論文執筆者への要望、及び今後の機関データ整備に関する NISTEP の考え方の 2 点について触れる。

2　データベースを用いたミクロデータ分析の現状と諸問題

2-1　分析のためのデータ源

(1) 全般的なデータ源

前章で、研究開発状況を把握するためのデータについて述べたが、これらのデータはどこから得られるであろうか。

○　インプットに関するデータ

インプットに関する研究者数、研究費のデータは、公表された統計から得られるが、それらはマクロレベルかせいぜいメゾレベル（大学部門、公的機関部門、民間部門といったセクターレベル）のデータに限られていた。しかし、最近、統計の個票データを研究目的で使用できるようになり、ミクロレベル分析が可能になりつつある。また、研究者ディレクトリ（我が国の場合、代表的なものとして Researchmap がある）を機関別に集計すれば、研究者数についてのミクロデータが得られる。特定の研究資金に関するミクロデータは、たとえば国立情報学研究所(NII)で作成・提供している科研費データベース(KAKEN)から得ることができる。

○　アウトプットに関するデータ

アウトプットに関するデータについては、論文や特許のデータベースが主要なデータ源となる。これらのデータベースから論文数や特許数のデータが得られるのは当然であるが、一部のデータベースには、論文の影響度の指標となる引用文献情報も含まれている。更に、最近、論文の謝辞に含まれる研究資金源の情報を収録するデータベースもあり、この情報はインプット分析にも利用できる。これらのデータベースには、論文の著者や特許の発明者(あるいは出願人)の情報はもちろん、多くの場合その所属機関の情報も含まれているので、ミクロデータ分析が可能である。

○　その他のデータ

研究者の研究交流や社会活動のデータはなかなか得にくいが、論文データベースに含まれる共著論文から、機関間、国際間の共同研究の状況や、共同研究に基づく研究者のネットワーク構造を分析することができる。

(2) 研究アウトプットのデータ源としての論文データベース

前述のように、研究開発状況の分析においては、論文や特許のデータベースが最も重要なデータ源であると考えられる。特に、大学や公的機関を対象にする場合は、研究アウトプットの中心は論文であるため、論文データベースの比重が増す。

全世界の主要な論文を収録するデータベースには、以下のようなものがある。

図表 3　　全世界の主要な論文を収録するデータベースの例

対象分野		論文データベース名	データベース提供機関
広分野を対象		Web of Science (WoS)	Thomson Reuters
		Scopus	Elsevier
		JSTplus	科学技術振興機構
特定分野を対象	化学	Chemical Abstracts (CA)	American Chemical Society
	医学	MEDLINE	National Library of Medicine
	医学	EMBASE	Elsevier
	物理・電気・情報	INSPEC	Institution of Engineering and Technology
	生物	BIOSIS	Thomson Reuters

これらのデータベースは、冊子体の時代から数えると数十年～百数十年の歴史を持ち、毎年数十万～百万件の論文データを追加している。

論文データベースに含まれる項目内容と含まれる情報を**図表 4** に示す。データベースの単位となるレコードは個々の論文である。**図表 4** の(a)、(b)、(c)に示す項目は、多くのデータベースに共通に含まれる(雑誌論文の場合であり、会議録論文やレポートの場合はやや異なる)。ミクロデータ分析では、(c)の著者に関する項目が重要である。

一方、(d)に示すように、データベースにより独自の項目も存在する。また、共通の項目であっても、データの表記法や表記規則はデータベースごとの特徴がある。

図表 4　論文データベースに含まれる項目内容と含まれる情報

	項目の種類	項目内容	含まれる情報
(a)	共通項目	レコード(論文)を識別するための項目	記事ID、DOI、出典情報(雑誌名、雑誌識別番号、発行年、論文掲載の巻号ページ)
(b)	共通項目	主題内容を示す項目	論文タイトル、抄録、主題索引語、主題分類
(c)	共通項目	著者に関する項目	著者名、著者所属機関、所属機関のアドレス
(d)	独自項目	特殊な索引データ項目	・引用索引(WoS、Scopus、CA) ・化学物質索引(CA) ・物質データ索引(INSPEC)　等

論文データベースにおけるデータ記述（データ項目とその表記）を**図表 5** により説明する。**図表 5** は、2007 年の Cell 誌に発表された山中伸弥博士の代表的論文を示す WoS の検索結果である。

(a) レコード（論文）を識別するための項目

「出版物名」の行に、この論文が、Cell の第 131 巻第 5 号（2007 年 11 月 30 日発行）の 861～872 ページに発表されたことが示されている。同じ行に、この論文の一意的識別子であるDOIも示されている。また、アクセッション番号は、このレコードに与えられた WoS の識別番号である。（**図表 5** にて、黄色のハイライト部分である。）

(b) 主題内容を示す項目

タイトル、抄録、KeyWords Plus、Web of Science の分野の各項目がこれに当たる。「Web of Science の分野」と「研究分野」には、Cell という雑誌に付与されている主題カテゴリーが示されている。（**図表 5** にて、水色のハイライト部分である。）

(c) 著者に関する項目

「著者名」には、この論文の 7 人の著者が並記されており、「著者所属」にはこれらの著者のそれぞれが所属する機関・組織が示されている。たとえば山中博士は、ここに挙げられた４つの組織のすべてに所属していることが判る。また、別刷り請求先とその E-mail アドレスの情報もある。（**図表 5** にて、桃色のハイライト部分である。）

(d) 特殊な索引データ項目

WoS や Scopus は、(a)～(c)に挙げた論文データベースの一般的項目の他に、引用文献の情報を含むことが特徴である。**図表 5** では、この論文には 30 の文献が引用（参照）されていること、4,505 の文献から引用されている（この論文を検索した 2014 年 1 月時点で）ことが判るが、検索画面でこれらの箇所をクリックすれば、引用文献や被引用文献のリストを見ることができる。なお、被引用文献は、WoS の全収録論文の引用文献を再編することにより得られる。（**図表 5** にて、緑色のハイライト部分である。）

図表 5 論文データベースのデータ例（Web of Science から検索）

レコード 1 / 1

タイトル: Induction of pluripotent stem cells from adult human fibroblasts by defined factors

著者名: Takahashi, K (Takahashi, Kazutoshi); Tanabe, K (Tanabe, Koji); Ohnuki, M (Ohnuki, Mari); Narita, M (Narita, Megumi); Ichisaka, T (Ichisaka, Tomoko); Tomoda, K (Tomoda, Kiichiro); Yamanaka, S (Yamanaka, Shinya)

出版物名: CELL **巻:** 131 **号:** 5 **ページ:** 861-872 **DOI:** 10.1016/j.cell.2007.11.019 **発行:** NOV 30 2007

Web of Science の被引用数: 4505

合計被引用数: 4869

引用文献数: 30

抄録: Successful reprogramming of differentiated human somatic cells into a pluripotent state would allow creation of patient- and disease-specific stem cells. We previously reported generation of induced pluripotent stem (iPS) cells, capable of germline transmission, from mouse somatic cells by transduction of four defined transcription factors. Here, we demonstrate the generation of iPS cells from adult human dermal fibroblasts with the same four factors: Oct3/4, Sox2, Klf4, and c-Myc. Human iPS cells were similar to human embryonic stem (ES) cells in morphology, proliferation, surface antigens, gene expression, epigenetic status of pluripotent cell-specific genes, and telomerase activity. Furthermore, these cells could differentiate into cell types of the three germ layers in vitro and in teratomas. These findings demonstrate that iPS cells can be generated from adult human fibroblasts.

アクセッション番号: WOS:000251800900010

言語: English

ドキュメントタイプ: Article

KeyWords Plus: SELF-RENEWAL; HUMAN BLASTOCYSTS; MOUSE EMBRYOS; ES CELLS; LINES; DIFFERENTIATION; ACTIVATION; STAT3; GENERATION; MAINTAIN

著者所属: [Takahashi, Kazutoshi; Tanabe, Koji; Ohnuki, Mari; Narita, Megumi; Ichisaka, Tomoko; Yamanaka, Shinya] Kyoto Univ, Inst Frontier Med Sci, Dept Stem Cell Biol, Kyoto 6068507, Japan.
[Narita, Megumi; Ichisaka, Tomoko; Yamanaka, Shinya] Japan Sci & Technol Agncy, CREST, Kawagoe, Saitama 3320012, Japan.
[Tomoda, Kiichiro; Yamanaka, Shinya] Gladstone Inst Cardiovasc Dis, San Francisco, CA 94158 USA.
[Yamanaka, Shinya] Kyoto Univ, Inst Integrat Cell Mat Sci, Kyoto 6068507, Japan.

別刷り請求先: Yamanaka, S（別刷り著者）、Kyoto Univ, Inst Frontier Med Sci, Dept Stem Cell Biol, Kyoto 6068507, Japan.

Email アドレス: yamanaka@ ac.jp

発行者: CELL PRESS

発行者所属: 600 TECHNOLOGY SQUARE, 5TH FLOOR, CAMBRIDGE, MA 02139 USA

Web of Science の分野: Biochemistry & Molecular Biology; Cell Biology

研究分野: Biochemistry & Molecular Biology; Cell Biology

IDS 番号: 243MG

ISSN: 0092-8674

29文字の出版物名略称: CELL

ISO 出版物名略称: Cell

ジャーナル項目ページ数: 12

（注1）トムソン・ロイター Web of Science （2014 年 1 月時点）の検索結果である。

（注 2）Email アドレスは一部加工している。

2-2 ミクロデータ分析の難しさ

データベースを用いて研究開発アウトプットの分析を行うには、**図表 6** の手順によるのが一般的である。

図表 6 論文データベースを用いて研究開発アウトプットの分析の流れ

大まかな分析手順		補足説明
①	データ源とするデータベースを選択する。	複数のデータベースを用いることもある。
②	分析の目的に応じて、データベースから分析対象とする論文を検索する。	たとえば、対象とする主題やテーマ、年代、国や地域等を指定して検索する。
③	検索されたデータをダウンロードする。	ほとんどのデータベース検索システムでは、定型的フォーマットによるデータのダウンロードを行う機能を備えている。
④	ダウンロードしたデータから、分析に不要なレコード（混入したノイズ等）を削除する。	
⑤	分析の前処理として、データの整理、クリーニングを行う。	
⑥	いろいろなデータ項目に関する集計、クロス集計、分類、その他の統計処理により分析を実行する。	

しかし、この手順によって研究開発の実態や動向を分析するのは実は容易なことではない。その最大の理由は、⑤のデータの整理、クリーニングに多大の労力を要することであり、全工程の過半の工数がこれに充てられると言っても過言ではない。逆に言えば、この過程をおろそかにしたデータ分析の精度、信頼性は低い。

この過程が重要であるのは、そもそも論文データベースは情報検索の目的で作られているため、データ分析に適した主題分類や、著者及びその所属機関の明確な識別がなされていないことに主な原因がある（データ分析のためのデータベース利用が進んでこの点での改善がなされているのは事実であるが）。

以下に、「論文の主題やテーマを分析する際の諸問題」、「著者を分析する際の諸問題」、「著者所属機関を分析する際の諸問題」を示す。データの整理、クリーニングの要点を示すが、特にミクロデータ分析で重要な著者所属機関データについて詳述する。

なお、データベースを用いて研究開発動向の分析を行う際には、この他に論文の収録方針や収録範囲による問題が重要であるが、これについてはここでは触れない。

（1）論文の主題やテーマを分析する際の諸問題

論文のタイトルや抄録、付与された索引語(キーワード)、主題分類等がこれに当たる。しかし、タイトルや抄録中の用語は、(i)主題に無関係な語が多く含まれる、(ii)同一の概念に対し多くの同義語、類義語が存在する、(iii)一つの論文には多面的な主題が含まれる、等の問題があり、主題による分類や集計を困難にする。シソーラス等に基づく統制索引語を分析に用いれば、(i)と(ii)の問題はほぼ解決されるが、(iii)の問題は相変わらず存在する。このため、用語や論文の類型化には、それぞれの分析の目的に応じて、用語の標準化、統合、カテゴリー化等の作業が必要とされる。このために因子分析やクラスター分析等の多変量解析の手法が用いられることもある。

主題分類は、用語に比べると包括的、体系的なので主題の分析に適しているが、分類の体系が分析の目的に適しているとは限らない。また、WoS や Scopus では、論文単位でなく雑誌単位に分類が付けられ、一つの雑誌に一般に複数の分類が付けられていることも、データ分析の観点からは不都合なことが多い。

このように、主題からの分析には多くの課題があるが、本稿ではこれ以上は触れない。

（2）著者を分析する際の諸問題

論文に示された著者名からその人物を同定することを「著者名寄せ」という。

著者名寄せで最も厄介なのは同姓同名の異著者の存在である。データベースによっては、著者名をフルネームでなく姓(last name)と名(first name, middle name)のイニシアルで表記するものがあり、この場合問題は更に深刻になる。同姓同名（あるいは同姓同イニシアル）の別著者を識別するための最も一般的な情報は論文の主題（分類や発表の雑誌の類似性）と所属機関であるが、異なる著者が同一機関に所属したり類似の分野の研究を行ったりすることもあるし、逆に同じ著者が所属や研究分野を変えることもある。共著者や引用文献の情報は著者名寄せに有効であることが示されているが、大量の論文データにこれを適用するには相当の労力が必要である。電子メールアドレスが一致すれば同じ著者と見てほぼ間違いないが、この情報がデータベースに含まれるとは限らない。現在のところ、これらの情報を組み合わせて、同姓同イニシアル著者の論文集合をクラスター分析等により類別することが、著者名寄せを正確に（完全ではないが）行う方法である。

同姓同名の別著者の識別と逆の問題として、結婚その他の理由による同一人物の異名の問題がある。しかし、個別の履歴情報を知らない限り、機械的にこれに対処することは難しい。

現在、研究者に一意の識別番号を付与することを目的とした ORCID （Open Researcher and Contributor ID）と呼ばれる国際プロジェクトが、学術出版、データベース構築、その他の機関の協力によって進められている。著者名寄せ問題の最終的解決は、この仕組みに世界の大多数の研究者が登録して識別番号を取得し、論文等にそれを表記することであろう。

著者データの分析に関して、名寄せ以外の留意すべき問題として、共著論文の計数法がある。論文に複数の著者がある場合、通常はそれらの著者それぞれが 1 本の論文を発表したとして、それを基に各研究者の論文数を数えることが多い。これを整数カウントという。しかし、単独で執筆した論文も 10 人の著者との

共著で執筆した論文も同じ 1 本と数えるのは不合理とも考えられる。また、この方法では、著者別に集計した論文数の合計が全論文数と合わないという統計操作上の問題がある。これらを避けるため、1 論文の各著者に寄与を分配し、全体で寄与が 1 になるような計数法がある。これを分数カウントという。著者が n 人であるとき寄与は 1/n ずつとするのが最も単純であるが、第一著者に高い配分を与える等、寄与に応じて不均等に配分する方法もある。最も極端なのは、第一著者のみに寄与 1 を与え、他の著者への配分は 0 とする方法である。データ分析でどちらの計数法を用いるかは、分析の目的や必要な作業量を勘案して決めるべきである。

(3) 著者所属機関を分析する際の諸問題

所属機関データのクリーニングでの最も重要な問題は、著者データの場合と同様に名寄せ(論文に表記されている機関名からその機関を同定すること)であるが、その内容は著者の場合とは異なる。それは、同じ機関の名称が論文により様々に表記されること、いわゆる「表記のゆれ」による問題である。従って、ある機関に属する著者の論文を検索あるいは同定しようとすれば、様々の表記ゆれを考慮しなければならない。

このような表記ゆれが起こる主な理由は、雑誌等に発表されたもとの論文で、著者により、あるいは雑誌により所属機関の表記が異なることによる。後述するように、データベース提供機関では、正確で簡便な機関検索のため様々な努力を行っているが、十分な解決には至っていない。

一つの例として、東京農工大学に所属する同じ著者が同じ雑誌に発表した 3 つの論文で、Scopus における所属機関名が次のようにそれぞれ異なるものがある((a)が正式の英語名称である)。

(a) Tokyo University of Agriculture and Technology
(b) Tokyo Noko University of Technology
(c) Tokyo A and T University

このような機関名の表記ゆれは、次の 6 つのパターンに類別化される。4－3(2)に具体的事例を示す。

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
③ 機関の略称
④ 冠詞、前置詞、接続詞の省略や書き換え
⑤ 機関の旧名の表記
⑥ スペルの違い(単数形と複数形のゆれ、ヘボン式と訓令式のゆれ等)

これらは典型的な(いわば狭義の)表記ゆれのタイプであるが、より広義にとらえると、以下のようなことも、機関名寄せを行うとき厄介な問題となる。

(a) 下部組織名の表記

機関レベルより深い組織(大学の学部や各機関の付属施設など)のレベルでミクロデータ分析を行いたい場合はしばしばあるが、これは機関レベルの分析より一層厄介である。なぜなら、機関の下部組織表記

では、上記の①～⑥に挙げたゆれが機関名表記より更に多様である以外に、下部組織に特有の別の問題があるからである。具体的には4－3(3)で述べる。

(b) 機関、組織の変遷

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。ミクロデータ分析では、このような場合、変遷前後の機関を関係づけたいことが多いので、変遷情報を把握することが重要である。

下部組織の変遷は更に甚だしく、大学、公的機関、民間企業を問わず、日常的に新設、統合、改組等が行われている。(a)で述べたように、ただでさえ下部組織の構成は複雑でその表記は多様である上、このように変遷が激しいので、下部組織レベルのミクロデータ分析は(必要性が高いにも拘わらず)極めて困難である。

(c) 同一著者の複数機関所属

論文の著者が複数の機関に所属していることがある。このような場合、論文の著者所属機関には複数の機関が併記されるのが通例なので、データベースにおいても、**図表 5** の山中博士の例にあるように、別々の機関データとして収録される。しかし、論文に著者が複数の機関(または組織)を合体して記述していると、データベースでもそれらが分離されず、一つの機関データに2機関が存在してしまう場合がある。次の2つの例はいずれも Scopus からとられたもので、(i)は東京大学物性研究所と科学技術振興機構(JST)の CREST が、(ii)は国立遺伝学研究所と総合研究大学院大学が合体表記された例である。ミクロデータ分析では、このような表記例もあることを考慮しなければならない。

(i) Institute for Solid State Physics, University of Tokyo, JST-CREST, Japan
(ii) Division of Mammalian Development, National institute of Genetics and Department of Genetics, SOKENDAI

以上のように多様な表記ゆれがあると、名称が類似した機関の間で誤同定(本来機関 A に対する表記を別の機関 B に同定)が生ずることに注意しなければならない。

この節に述べた機関名表記ゆれの実態や、表記ゆれによって誤同定が起こる可能性については、NISTEP の経験に基づき、第4章で具体的に述べる。

2-3 データベース提供機関による検索・同定の簡易化に関する動き

前節で述べたような同名異人著者の識別、多様な表記ゆれが存在する機関の同定を少しでも容易にするため、データベース提供機関ではいろいろな努力をしている。そのいくつかについて簡単に触れる。但し、その手法や仕組みは必ずしも公表されていないので、それぞれの試みがどのデータベースで行われているかについては記さない。

(1) 著者の識別について

(a) 著者フルネームの記載

著者名について姓とイニシアルのみで表記していたデータベースが、フルネームを記載するようになった。

(b) 著者と所属機関の対応付け

著者とその所属機関のデータが独立であったデータベースにおいて、その対応付けが行われるようになった。

(c) 著者の属性を示すデータの増強

情報が得られた場合、著者の電子メールアドレスや研究者識別番号(たとえば前述の ORCID により与えられる ID)を記載する。

(d) データベース独自の著者 ID の付与

データベース側で、論文に付属したいろいろなデータを用いて同一著者と推定される論文を同定し、それらの著者に識別番号を与える。更に、研究者からの申請により識別の改善を行う。

(2) 機関の同定について

(a) 機関名表記の統一

それぞれのデータベースにおいて、機関名表記のためのできるだけ統一的な方法や基準を定める。多くの場合、名寄せのための辞書やアルゴリズムにより、統一的表記に変換する。

(b) データベース独自の機関 ID の付与

データベース中の個々の機関表記に対して、名寄せの結果同定された機関識別番号を付与する。検索時に利用者が入力した機関名から、その表記に対応する機関を判定し、その識別番号を持つ論文を回答する。

3　大学・公的機関に関するデータ整備－NISTEP における取組み

第 2 章の記述から、研究開発におけるミクロデータ分析を正確かつ高精度に行うには、データベースにおける機関表記ゆれの実態の把握と、それに基づく名寄せのための手法やツールの開発、重要性がご理解いただけたと思う。データベースから得られる研究アウトプットデータと、他の情報源から得られる研究インプット等のデータの有効な接続も、これなしには行えない。しかし、ミクロデータ分析を行おうとする個人や機関がこれに対処するのは容易なことではない。2－3で述べたように、データベース提供機関の努力により改善も見られるが、まだ道半ばである。この章では、これに関する NISTEP の活動について述べる。

3-1 「大学・公的機関に関するデータ整備」の概要

NISTEP では、文部科学省の「科学技術イノベーションにおける“政策のための科学”推進事業」の一環として、平成 23(2011)年度から「データ・情報基盤の構築」を実施している（これについては参考文献[1]～[6]を参照されたい）。その中のプロジェクトの一つが、「大学・公的機関に関するデータ整備」である。これは、我が国における研究開発（特に政府予算で実施されているもの）の実態の把握・分析及びそのパフォーマンス評価を、国、セクター、個別機関などの各レベルで行うための基礎として、大学・公的機関の科学技術生産に関するデータの整備を行うことを目的としている。このため、研究開発統計と科学論文のデータベース、さらに部分的に特許のデータベースを用い、それらを様々なレベルで整備して、相互にデータ接続をしようとする計画である。

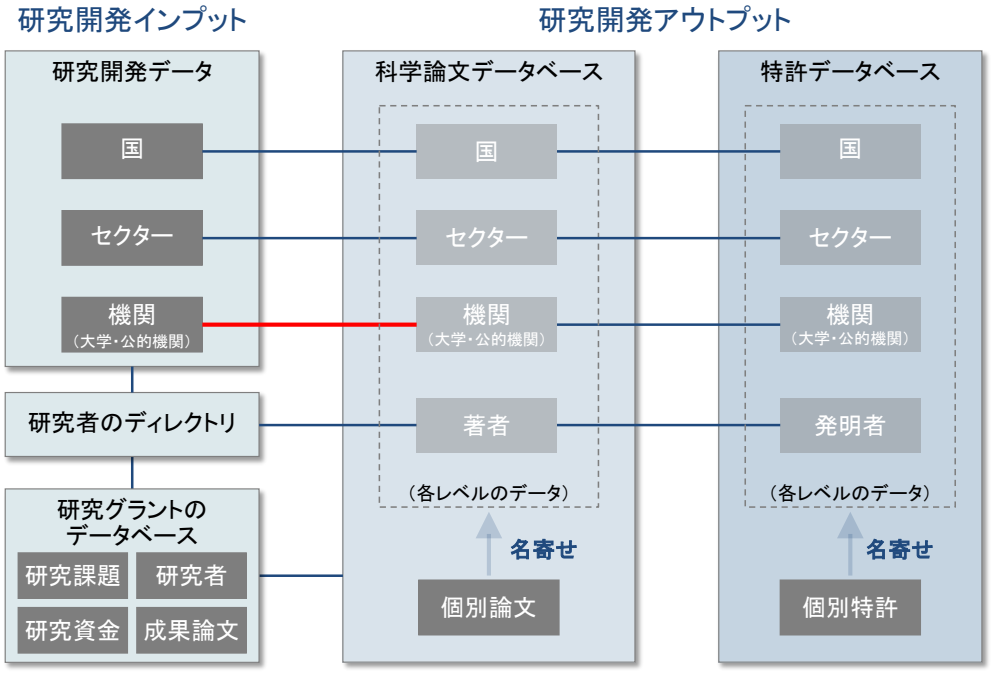
図表 7 はその構想をモデル的に示したものである。研究開発インプットと研究開発アウトプットのデータが、国レベル、セクターレベル、機関レベル、研究者レベルで接続されている。このうち国レベルのデータ接続には大きな問題はない。それぞれのデータベースから特定の国のデータを取り出すことは比較的容易だからである。しかし、他のレベルのデータ接続は単純ではない。

図表 7 左に示す研究開発インプットのデータ(研究者数、研究費等)については、研究開発統計から国レベル、セクターレベルのデータが得られるだけでなく、統計の個票データを研究目的で使用することにより、機関レベルデータの取得も可能になりつつある。しかし、研究開発アウトプットのデータ源である論文データベースや特許データベースから、機関レベル、セクターレベルのデータを得ることは簡単ではない。論文データベースについては、第2章で述べたように著者所属機関の表記ゆれが大きく、その名寄せが必要である。また、セクターレベルの集計には、各機関のセクターを示す辞書が必要である。特許データベースの出願人データに示される機関名には、論文データベースほどの表記ゆれはないが、セクター分類には同じ問題がある。また、研究者レベルの分析では、論文の著者データや特許の発明者データ中の個人名から同名異人や異名同人を識別するための名寄せが必要である。

このような名寄せによって、セクターレベル、機関レベル、研究者レベルの集計やインプットデータとの接続が行われる。また、謝辞データを含む論文データベースを、左下の研究グラントのデータベース（科研費研究テーマのデータベース等）と接続することにより、研究資金（ファンディング）データのミクロ分析も可能になる。

NISTEP の「データ・情報基盤の構築」プロジェクトは、ミクロデータ分析を行う研究者や政策担当者を支援する基盤データやツールを開発することを主な目的としている。

図表 7　大学・公的機関における研究開発に関するデータ整備の概念モデル



3-2 主要な整備データとその公開

この節では、当面この事業で中心的に取り組んでいる次の 2 つのデータ整備について述べる。

(1) NISTEP 大学・公的機関名辞書(以下単に「機関名辞書」という)の整備
(2) 論文データベースにおける機関名寄せ:機関名辞書とのミクロ接続

これらの整備データは、関係者に利用していただくため、NISTEP の「データ・情報基盤」の Web サイトから公開を進めている。

http://www.nistep.go.jp/research/scisip/randd-on-university

個々の公開データについては、以下のそれぞれの項で述べるが、ミクロデータ分析、その他日本の研究機関に関する分析に際し正確で高精度な機関同定を行うための活用が期待される。

(1) 機関名辞書の整備

機関名辞書は、インプットデータ、アウトプットデータを機関レベル及びセクターレベルで分析するための基本情報を含むもので、このプロジェクトの中核的役割を果たす。

この辞書に含まれるのは、研究開発を行っている日本国内の機関である。名称にあるように大学、公的機関を重点とするが、地方公共団体の機関、企業、非営利法人等もできるだけ含めており、全部で 10,000 機関以上に達している。それぞれの機関には、NISTEP 独自の識別 ID を与える。収録している情報を**図表 8** に示す。

また、**図表 9** には、この辞書で用いているセクター分類と、現在公開している機関名辞書に含まれる機関数を示す。

図表 8　NISTEP 大学・公的機関名辞書に収録されている基本情報

基本情報の種類		内容
①	NID	●NISTEPが独自に与える機関識別ID。
②	機関の名称	●日本語の正式名称、英語の正式名称に加え、英語の通称、略称もできるだけ収録。 ●各名称には、正式名称と確認したものとそれ以外を区別するフラグを与える。
③	セクター	●産・学・官よりかなり細かく、16のセクターに各機関を分類。この分類は、科学技術研究調査で用いられているものに近い。 ●これとは別に、病院の機関にはそのことを示すフラグを与える。
④	機関の下部組織	●研究活動に関して主要な機関について一部の下部組織も収録。特に、主要大学の学部・研究科・附置研究所等、大学共同利用機関である機構に属する各研究所、一部の独立行政法人に属する機関は網羅的に収録。 ●上部機関と下部組織の間に関係づけを行う。
⑤	機関の変遷情報	●この15年ほどの間に統廃合、改組、名称変更等があって現存しない機関についても、できるだけ収録。 ●変更のあった日付、継承機関（存在する場合）等の情報も収録。

図表 9　機関名辞書で使用するセクターとセクターごとの収録機関数

セクター	収録機関数	セクター	収録機関数
国立大学	101	私立高専	3
国立短大	26	大学共同利用機関	5
国立高専	59	国の機関	135
公立大学	94	特殊法人・独立行政法人	133
公立短大	62	地方公共団体の機関	696
公立高専	6	会社	4,421
私立大学	601	非営利団体	3,586
私立短大	515	その他の機関	6
		計	10,449

（注）現存しない機関を含み、下部組織を含まない。

このように、かなり細かいセクター分類が付与されているので、その観点からの分析が可能である。また、機関の変遷情報（統廃合や名称変更）が含まれており、変遷前後の機関 ID がリンク付けされているので、これもミクロデータ分析には有用である。たとえば、産業技術総合研究所の発表論文数の推移を、旧工業技術院の各研究所と併せて集計することができる。

機関名辞書は上記の「データ・情報基盤」サイトから公開している(Ver.2012.1)。もとの機関名辞書はリレーショナルデータベース型の構造であるが、公開版は、サブファイルをひとつのテーブルに統合した形の Excel ファイルである。その内容は**図表 8** に示すとおりであるが、2014 年度には、次の点を拡充した改訂版を公開の予定である。

① 機関の英語別名、略称の追加
② 下部組織を網羅的に収録する大学の増加（現在の 12 大学を 32 大学に）
③ 機関の変遷情報の充実

(2) 論文データベースにおける機関名寄せ:機関名辞書とのミクロ接続

2－2(3)で述べた機関名寄せが重要な役割を持つのがこの部分である。ここでは活動の内容を述べ、その過程で得られた表記ゆれデータについて次章で分析する。

(a) 対象としたデータ

世界的な論文データベースである Scopus と Web of Science Core Collection (以下 Web of Science または WoS)を対象とした。当面この 2 つを対象とした理由は次の通りである。

① 広い分野(科学技術分野のみならず人文・社会科学関係も含む)を対象としている。
② 引用文献データを含んでいる。
③ データ分析によく使われるデータベースであり、各国における分析、OECD のような国際機関での分析でも利用されている。。

これらのデータベースから、1996～2011 年の期間に発表された日本の論文(日本の機関に属する著者を少なくとも一人含む論文)を抽出し、そこに含まれる機関の名寄せ(機関名辞書の登録機関への対応づけ)を行った。

(b) 名寄せ作業

データベース中の個々の機関データを、機関名辞書中の名称データと照合することにより名寄せを行う。しかし、データベース中のデータには2－2(3)で述べたように様々な表記ゆれがあるため、名寄せアルゴリズムの開発は、結果から誤りや問題点を見出しては修正するという試行錯誤の繰り返しであった。

名寄せアルゴリズムの手順は次の通りである。

図表 10 名寄せアルゴリズムの手順

名寄せアルゴリズムの手順	
	整備されたNISTEP大学・公的機関名辞書を用意する。
①	データベース中の機関名データを、機関名辞書中の機関名データと最長マッチングを行い、マッチした機関に同定する。機関名辞書に下部組織も収録されていれば、それとのマッチングも行う。
②	①で同定できなかったデータについて、機関名辞書との曖昧マッチングを行い、マッチした機関に同定する。
③	②でも同定できなかったデータに郵便番号が含まれていれば、機関名辞書中の郵便番号データとマッチングを行わせ、マッチした機関に同定する。
④	機関同定ができなかったデータは、データに含まれる機関種別を表す文字列(″Co. LTD.″など)からセクター同定を行う。また、″Hospital″等が含まれていれば病院と同定する。
⑤	以上のいずれにも同定できないデータは同定不能とする。

(c) 名寄せの結果

1996-2011 年の期間において、Scopus では延べ 329 万件、WoS では延べ 278 万件の日本機関データが出現した。このうち機関同定できた(上記**図表 10** の①～③)のは、Scopus では 91.9%、WoS では 93.6%であった(2012 年度末時点)。機関同定はできなかったが④の同定ができたのは、それぞれ 5.4%、4.6%であった。機関同定できたデータのサンプリング調査により、同定の精度は 98%以上であることを確認した。

2013 年度は、これらの結果(特にエラー同定データや同定できなかったデータ)の評価に基づき機関名辞書と名寄せアルゴリズムの改善を行い、同定率と同定精度を更に向上させることができた。しかし一部課題も残っており、それについては 2014 年度に検討する予定である。

(d) データの公開

2012 年度末時点でのデータ整備に基づき、次のデータを、前述の「データ・情報基盤」サイトから公開している。

① 大学・公的機関名英語表記ゆれテーブル(Ver.2013.1)

ScopusとWoSにおける日本の大学・公的機関の表記ゆれを調査した結果の一部を、Scopusの提供元であるElsevier社とWoSの提供元であるトムソン・ロイターの了解を得て公開している。Scopus では、1996-2010 年の期間に延べ 1,000 以上出現した 205 の大学と 40 の公的機関について、10 回以上出現した 1,461 の表記バリエーションを記載している。また、WoS では、1996-2011 年の期間に延べ 800 以上出現した 219 の大学と 56 の公的機関について、8 回以上出現した表記バリエーション 968 を記載している。これによって、これらの機関のデータのほとんどをカバーし、国内全機関データの 60%以上をカバーするので、これらのデータベースで機関名検索を行う場合の有効な補助ツールになると考えられる。

② Scopus-NISTEP 大学・公的機関名辞書対応テーブル(Ver.2013.1)

機関名寄せの結果、論文データベースの著者所属機関情報(論文 ID と論文内機関番号の組み合わせ)と機関名辞書の機関 ID が対応づけられる。Elsevier 社の了解を得て Scopus についてのこの対応テーブルを公開している。但し、大学、公的機関以外の機関(地方公共団体の機関、会社、非営利法人等)については現状では同定の精度がやや低いので、公開版では機関 ID を示さずセクターのみを公表している。

Scopus の利用者は、このテーブルをたとえば次のように活用することができる。

- Scopus で検索した論文データ集合における所属機関を、このテーブルを用いて同定
- ある機関の論文の一括検索(その機関 ID を持つ Scopus 論文データの集合をこのテーブルを用いて作成)
- 機関別又はセクター別の論文生産統計の作成と分析

なお、WoS についてもこれと同様の対応テーブルを公開するため、準備を進めている。

3-3 その他のデータ整備活動

ここでは、「大学・公的機関に関するデータ整備」で実施しているその他の活動について述べる。これらについてはまだデータの公開に段階に至っていないが、データ整備の進捗、関係者のニーズ等を勘案して今後検討したい。

(1) 研究インプットデータベースと機関名辞書とのミクロ接続

研究インプットデータベースとして 2002～2011 年度の科学技術研究調査対象名簿(以下「科技調査名簿」という)を選び、これと機関名辞書の機関名を対応づけるためのプログラムを開発した。科技調査名簿は大学学部等の下部組織単位になっているので、機関名辞書で下部組織が登録されているものはその単位でマッチングが可能である。科技調査名簿と機関名辞書の名称表記の違い、旧漢字と新漢字の違い、漢字コードの微妙な違い等を考慮したマッチング処理が可能である。科技調査名簿には各機関の研究費、研究人員のデータが含まれているので、次に述べる論文生産統計のデータを組み合わせれば、各機関の研究インプットと研究アウトプットの関係を示すデータを得ることができる。

(2) 論文生産統計のためのテーブル設計

3－2(2)で述べたミクロ接続の結果に基づいて、1996～2012 年の ScopusとWoS のデータに対する種々の論文生産統計が得られる。次のような集計が可能なようにテーブル設計を行った。

(a) 機関別集計とセクター別集計

機関別集計では、下部組織を独立させた場合と上位機関にまとめた場合、及び非現存機関を継承機関に合体した場合が可能である。また、これとは別に病院のみの集計も可能とした。

(b) 年別、分野別、及びこの両者を組み合わせた集計

(a)の結果を年別、分野別、年・分野別に集計できる(但し、分野別集計はWoS のみ)。分野にはWoS の雑誌主題カテゴリーを用いた。

(c) 整数カウントによる集計と分数カウントによる集計

分数カウントの場合の寄与は、データベースで著者と所属機関が対応付けされている場合は著者数比例配分、そうでない場合は各機関に均等配分とした。

この他、WoS を対象として、共著のタイプ別集計(単独機関内、同一セクター内、異セクター間、国際間)、国内機関と共著の多い外国機関の集計、海外の主要大学の生産統計も可能である。

（3） 特許データベースと機関名辞書のミクロ接続

特許データベースの出願人データに含まれる大学・公的機関を抽出し、機関名辞書とマッチングさせる手法を検討した。

（4） 論文謝辞からの研究資金源の分析

WoSでは、2008年後半から、論文の謝辞(acknowledgments)に現れる研究資金受給情報を収録している。このデータを用いて、日本の研究資金提供機関・制度について試行的な集計・分析を行った。

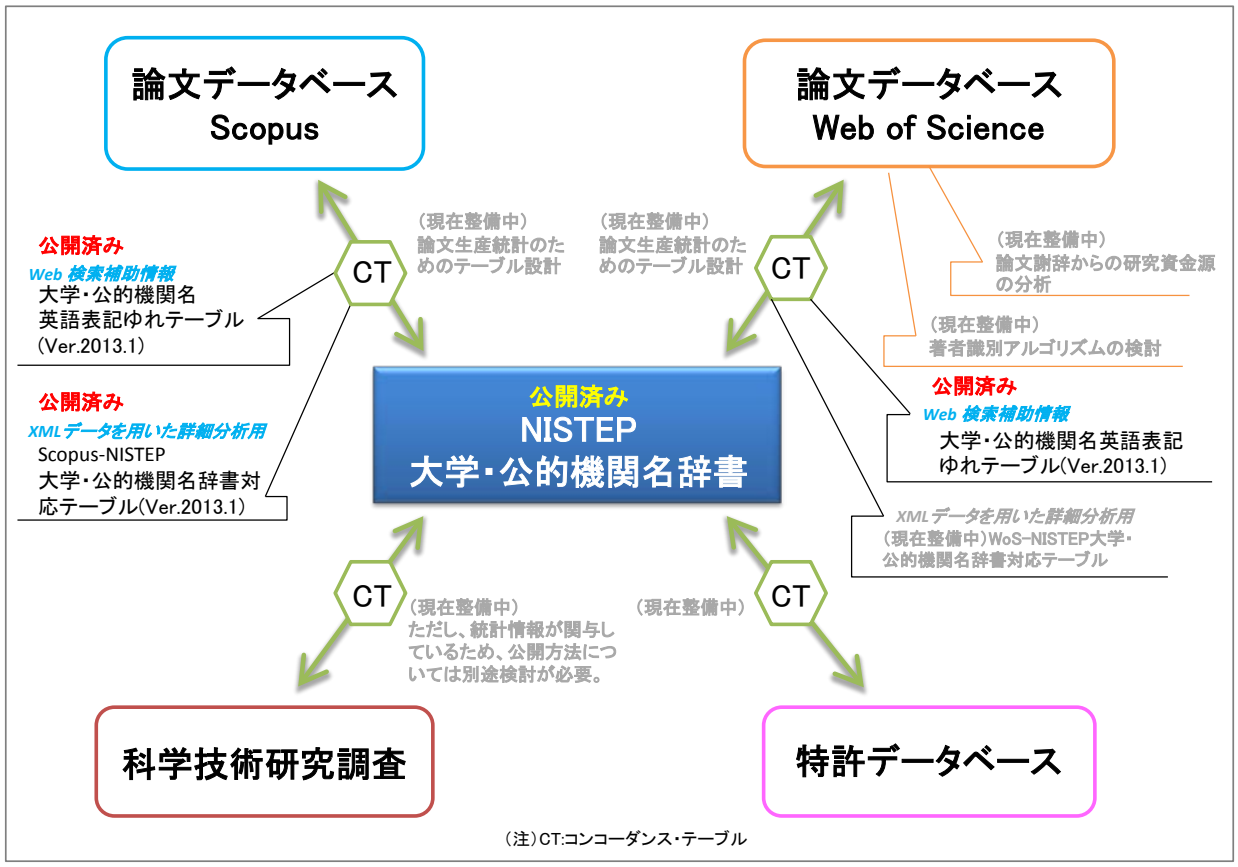
（5） 著者識別アルゴリズムの検討

WoS に含まれる日本の研究機関に所属している著者について、論文の諸データ（所属機関、研究分野、共著者等）を用いて著者名寄せ（同姓同イニシアル著者の識別）を行う手法を、試験的に検討した。

図表 11 は、機関名辞書を中心とした各種研究開発関連データの接続イメージである。機関名辞書を中心にして、多種多様な研究開発関連データが接続され、その結果、多様な変数の組合せによるデータ分析が可能となる。これは、より多くの仮説の検証の機会を提供することを意味する。

機関名辞書は、我が国の大学や公的機関を網羅的に収録していること、Web 上で公開して誰でも自由に利用できることが特徴である。利用をオープンにすることによって、様々な研究者が研究を行う際に、大学・公的機関に関する情報の典拠としての役割を果たすことを一つの目標としている。

図表 11　機関名辞書を中心とした各種研究開発関連データの接続



4 機関名表記ゆれの分析

3－2(2)で述べたデータベースにおける機関名寄せは、NISTEP の「大学・公的機関に関するデータ整備」事業の中で最も注力している作業であり、これは機関名辞書の充実とも深く結びついている。前述したように、名寄せの対象としたデータは、Scopus と WoS に収録された 1996～2011 年発表論文に含まれる日本の著者所属機関データである。この期間の日本の論文は、Scopus で 150 万件強、WoS で 140 万件弱であり、その中の国内機関データ数は、Scopus で延べ 329 万、WoS で延べ 278 万に及ぶ。

この章では、Scopus の 1996～2010 年のデータ(2011 年データは未整理)に基づき、機関名の表記ゆれデータの分析の一部を述べる。Scopus のデータを用いた理由は、WoS では University を Univ に置き換える等、機関名の表記にある程度の標準化がなされているのに対し、Scopus では原論文での表記がかなり保持されているからである。

4-1 分析の対象

Scopus からすべての表記ゆれバリエーションをとると膨大な数になるので、調査対象については下記の両方を満たす機関に限定した。

① 大学(短大、高専、大学共同利用機関を含む)または公的セクター(国の機関、特殊法人・独立行政法人)に所属する機関(地方自治体に属する機関、民間企業、非営利法人等は除く)
② 同定された所属機関データが 1,000 以上

そして、これらの機関に対して 10 回以上出現した表記バリエーションを分析した。このうち大学共同利用機関は、名称表記のパターンが大学よりも公的機関に似通っている。従って、以下では、大学共同利用機関は「公的機関」の範疇に含めることとする。

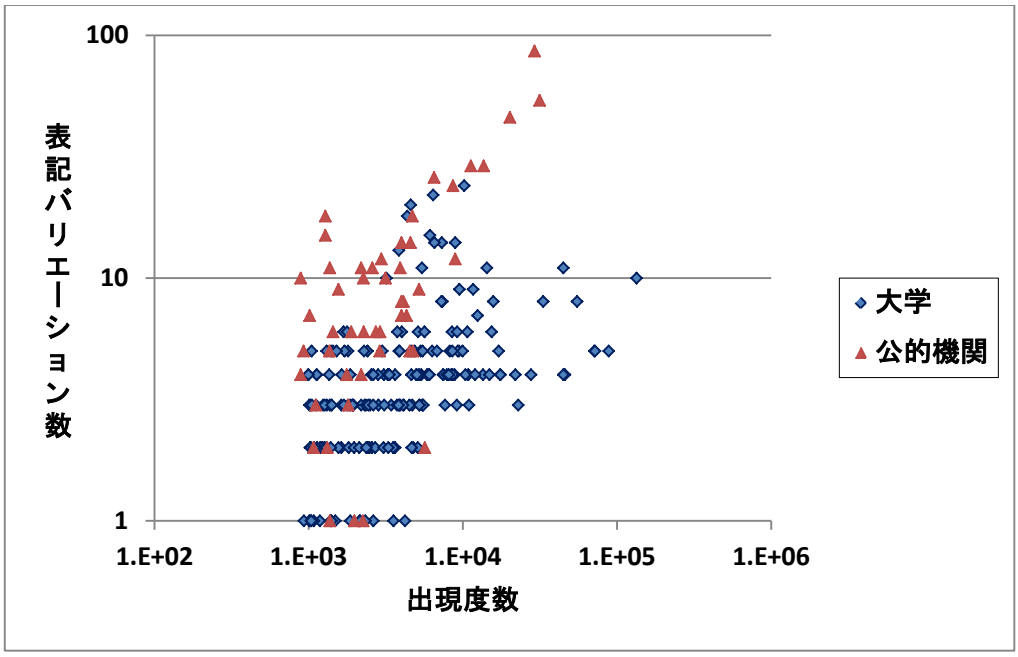
その結果、対象の機関は 245(大学 196、公的機関 49)、表記バリエーションの総数は 1,461(大学 853、公的機関 608)となった。これら出現度数 10 以上の表記バリエーションは、対象 245 機関の全データの 98～99%を、また Scopus の全国内機関データの約 63%をカバーする。

4-2 表記ゆれの分散の大きさ

上述のように、対象とした 245 機関の出現度数 10 以上の表記バリエーション数は 1,461 なので、1 機関平均ほぼ 6 のバリエーションがある。大学の平均 4.4 バリエーションに対して公的機関は平均 12.4 バリエーションで、後者の表記ゆれがずっと大きいが、この理由は後述する(4－4参照)。

機関の総出現度数が多ければ、当然表記バリエーション数も多くなると予想されるが、**図表 12** に示すように、確かにその傾向はあるものの相関は弱い(大学、公的機関とも、スピアマン順位相関係数は約 0.55)。機関によって様々な表記がなされるところと、比較的表記が安定しているところがある。

図表 12　Scopus における機関の出現度数と表記バリエーション数の関係



ある機関の表記ゆれの大小を判断するには、次の 2 通りの考え方がある。

(a)　表記バリエーションの数が多い。

(b)　バリエーション間に度数が均一に分散している。

そこで、次の 2 つの指標により、各機関の表記ゆれの程度を評価した。

[指標 A]エントロピー(H)

上記(a)と(b)の両方を考慮した指標であり、次式で表される。

$$H = \sum_{j=1}^{n} f_j \log_2\left(\frac{1}{f_j}\right)$$

ある機関がnの表記バリエーションを持つとき、f_jはj番目の表記バリエーションの相対度数(j= 1,2,…,n)である。機関の表記がn個のバリエーションに均一に分散していれば(f_1 = f_2 =…= 1/n) H = $\log_2$ n、完全に統一されていれば(f_1 = 1、f_2以下は 0) H = 0 となる。この指標は、全度数の大きい機関(大規模な機関)が大きくなる傾向がある。

[指標B]相対エントロピー(H_R)

これは上記の(b)のみを考慮した指標であり、次式で表される。

$$H_R = H \Big/ \log_2 n \qquad (n > 1)$$

$$H_R = 0 \qquad (n = 1)$$

機関の表記が n 個のバリエーションに均一に分散していれば H= 1 (nに無関係に) であり、ある一つのバリエーションが全度数に占める比率が大きくなるほど H= 0 に近づく。

ここでの計算は、出現度数 10 未満のバリエーションを考慮していないが、各機関の表記ゆれの程度を測る目安にはなると考えられる。

指標 A の値が 1.0 以上の機関は、大学では 191 中 28 (15%)であるのに対し公的機関では 49 中 34 (69%)であり、指標 B の値が 0.5 以上の大学は 21 (11%)、公的機関は 24 (49%)であった。これらの数字からも、公的機関の表記ゆれが大学よりずっと大きいことが判る。

4-3 大学における機関表記のゆれ

(1) 表記ゆれの程度が大きい大学

指標 A と指標 B の値が大きい機関を、それぞれ**図表 13** に示す。

図表 13 表記ゆれ指標の大きい大学

(a) 指標Aの値が大きい大学

順位	大学名	表記数	指標A値
1	総合研究大学院大学	18	2.56
2	北陸先端科学技術大学院大学	20	2.47
3	産業医科大学	22	2.32
4	東京薬科大学	10	2.25
5	富山医科薬科大学	11	2.17
6	聖マリアンナ医科大学	13	2.15
7	奈良先端科学技術大学院大学	15	1.83
8	東京農工大学	24	1.68
9	大阪電気通信大学	5	1.60
10	北海道医療大学	6	1.59

(b) 指標Bの値が大きい大学

順位	大学名	表記数	指標B値
1	福井医科大学	2	0.90
2	大阪薬科大学	3	0.80
3	香川医科大学	2	0.71
4	神戸学院大学	3	0.70
5	大阪電気通信大学	5	0.69
6	東京薬科大学	10	0.68
7	滋賀県立大学	2	0.66
8	富山医科薬科大学	11	0.63
9	明治薬科大学	2	0.63
10	北海道医療大学	6	0.62

(注) Elsevier Scopus を基に科学技術・学術政策研究所が集計

指標 A の大きい大学のうち、表記バリエーションが多い産業医科大学の例を**図表 14** に、東京農工大学の例を**図表 15** に示す。いずれも、リストの最初にある最も出現頻度の高い表記が正式の英語名であるが、出現度数 10 以上の全表記の中でのその出現率は、産業医科大では 47%、東京農工大では 60%に留まる。

次に、指標 B の値が大きい大学から、福井医科大学(福井大学と統合して現在は存在しない)の例を**図表 16** に、東京薬科大学の例を**図表 17** に示す。福井医科大学は、出現度数 10 以上の表記バリエーションは 2 つしかないが、2 番目の表記もかなりの出現比を占めるため、指標 B の値が高くなっている。東京薬科大学は多くの表記に分散しているが、最初の 2 つの表記がほぼ均等の度数を持つため、この指標値が高くなる。

図表 14　産業医科大学の表記ゆれ

機関名	機関ID	セクター	Scopusにおける表記ゆれ	英語正式名	出現度数
産業医科大学	NID201200202372859	私立大学	University of Occupational and Environmental Health	○	3010
産業医科大学	NID201200202372859	私立大学	Univ. of Occup. and Environ. Health		1876
産業医科大学	NID201200202372859	私立大学	Univ. of Occup./Environmental Health		450
産業医科大学	NID201200202372859	私立大学	University of Occupational and Environmental Health, Japan		297
産業医科大学	NID201200202372859	私立大学	Univ. Occup. Environ. Hlth., Japan		237
産業医科大学	NID201200202372859	私立大学	Univ. of Occup./Environ. Health		79
産業医科大学	NID201200202372859	私立大学	Univ. of Occupational/Envtl. Hlth.		59
産業医科大学	NID201200202372859	私立大学	Univ. of Occupational/Environ. Hlth.		57
産業医科大学	NID201200202372859	私立大学	Univ. Occup./Environ. Hlth., Japan		56
産業医科大学	NID201200202372859	私立大学	University of Occupational and Environmental Health Japan		48
産業医科大学	NID201200202372859	私立大学	Univ. of Occupational/Envtl. Health		47
産業医科大学	NID201200202372859	私立大学	Univ. of Occup. and Environ. Hlth.		37
産業医科大学	NID201200202372859	私立大学	Univ. Occup. Environ. Hlth. Japan		29
産業医科大学	NID201200202372859	私立大学	Univ. of Occup./Environ. Hlth. Japan		25
産業医科大学	NID201200202372859	私立大学	Univ. Occup./Environ. Health, Japan		21
産業医科大学	NID201200202372859	私立大学	Univ. of Occup. Environmental Health		18
産業医科大学	NID201200202372859	私立大学	University of Occupational and Environmental Health (UOEH)		17
産業医科大学	NID201200202372859	私立大学	University of Occupational Environmental Health		15
産業医科大学	NID201200202372859	私立大学	Univ. of Occup. and Envtl. Hlth.		14
産業医科大学	NID201200202372859	私立大学	Univ. of Occup. and Environ. H.		13
産業医科大学	NID201200202372859	私立大学	Univ. of Occupational and Environmental Health		13
産業医科大学	NID201200202372859	私立大学	UOEH		10

(注) Elsevier　Scopus を基に科学技術・学術政策研究所が集計

図表 15　東京農工大学の表記ゆれ

機関名	機関ID	セクター	Scopusにおける表記ゆれ	英語正式名	出現度数
東京農工大学	NID201200980805842	国立大学	Tokyo University of Agriculture and Technology	○	6166
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agric. and Technology		2993
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agric. and Technol.		227
東京農工大学	NID201200980805842	国立大学	Tokyo Noko University		121
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agriculture/Technol.		112
東京農工大学	NID201200980805842	国立大学	Tokyo University of Agriculture and Technology (TUAT)		109
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agric./Technology		79
東京農工大学	NID201200980805842	国立大学	Tokyo University of A and T		67
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. Agric. T.		53
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agriculture and Technology		40
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agri. and Tech.		33
東京農工大学	NID201200980805842	国立大学	Tokyo University of Agric./Technol.		33
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agriculture Technol.		25
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agriculture/Tech.		17
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agric. and T.		15
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agr. and Tech.		14
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agriculture and Tech.		14
東京農工大学	NID201200980805842	国立大学	Tokyo University of Agri. and Tech.		14
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of Agri. and Technology		13
東京農工大学	NID201200980805842	国立大学	University of Agriculture and Technology		13
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. Agric. and Technology		12
東京農工大学	NID201200980805842	国立大学	Tokyo University of Agriculture and Technology (TAT)		11
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. Agriculture/Technology		10
東京農工大学	NID201200980805842	国立大学	Tokyo Univ. of A and T		10

(注) Elsevier　Scopus を基に科学技術・学術政策研究所が集計

図表 16　福井医科大学の表記ゆれ

機関名	機関ID	セクター	Scopusにおける表記ゆれ	英語正式名	出現度数
福井医科大学	NID201200228019057	国立大学	Fukui Medical University	○	1658
福井医科大学	NID201200228019057	国立大学	Fukui Medical School		762

(注) 2003 年に福井大学に統合されている。

(注) Elsevier　Scopus を基に科学技術・学術政策研究所が集計

図表 17　東京薬科大学の表記ゆれ

機関名	機関ID	セクター	Scopusにおける表記ゆれ	英語正式名	出現度数
東京薬科大学	NID201200689092004	私立大学	Tokyo University of Pharmacy and Life Science	○	1030
東京薬科大学	NID201200689092004	私立大学	Tokyo Univ. of Pharm. and Life Sci.		1106
東京薬科大学	NID201200689092004	私立大学	Tokyo University of Pharmacy and Life Sciences		572
東京薬科大学	NID201200689092004	私立大学	Tokyo Univ. of Pharm./Life Science		248
東京薬科大学	NID201200689092004	私立大学	Tokyo Univ. of Pharmacy/Life Science		90
東京薬科大学	NID201200689092004	私立大学	Tokyo College of Pharmacy		62
東京薬科大学	NID201200689092004	私立大学	Tokyo Univ. Pharm. Life S.		40
東京薬科大学	NID201200689092004	私立大学	Tokyo Univ. of Pharm. and Life S.		33
東京薬科大学	NID201200689092004	私立大学	Tokyo Univ. of Pharmacy/Life Sci.		15
東京薬科大学	NID201200689092004	私立大学	Tokyo Univ. of Pharm./Life Sci.		12

(注) Elsevier　Scopus を基に科学技術・学術政策研究所が集計

（2）大学の機関名の表記ゆれのパターン

２－２(3)で、機関名の表記ゆれを６つのタイプに類別化したが、機関の名寄せを行う過程で認められた機関名の表記ゆれについては、次のようなパターンがある。ここでは、Scopus での実例を挙げて説明する。

[1] 正式の名称とは単語や語順が異なる表記

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。**図表 15** の Tokyo Noko University もこのタイプの表記である。

この種の表記ゆれで非常に難航するのが、静岡大学(Shizuoka University)と静岡県立大学(University of Shizuoka)である。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(422)、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。大学ホームページを用いて、個人ごとの特定も行っているが、その中では大学の研究者自体が間違っていることが認められる。

医科大学では、**図表 16** の福井医科大学の例のように、Medical University、Medical School、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。自治医科大学も同様である。また、東京慈恵会医科大学は Jikei University と Jikei University School of Medicine が、獨協医科大学は Dokkyo University School of Medicine と Dokkyo Medical University が、ともに公称とされているようで、それぞれが相当の頻度で使われている。

[2] 単語の略記

University、Institute、Medicine (Medical)、Science (Scientific)をそれぞれ Univ、Inst、Med、Sci とするような表記である。細かいことだが、略記語の末尾にピリオドを付けるか付けないかでも異なる表記バリエーションになる。

図表 14 や**図表 15** では、この種の表記ゆれが特に多い。東京農工大学の場合、“Agriculture and Technology”を“Agric. and T.”や“A and T”と略する極端な例も見られる。

[3] 機関の略称

図表 14 の産業医科大学のリストには、略称である UOEH や、通常の表記の後にこの略称をカッコに入れた表記がある。総合研究大学院大学(The Graduate University for Advanced Studies)は、GUAS とも SOKENDAI とも略記される。

[4] 冠詞、前置詞、接続詞の省略や書き換え

冠詞では、University で始まる大学名（University of Tokyo、University of Tsukuba など）の先頭に“The”を付ける表記と付けない表記が存在する。“of”、“for”、“and”等の前置詞や接続詞が省略されて表記されることも多い。また、上記リスト中の産業医科大学、東京農工大学、東京薬科大学の例に見られるように、“and”の代わりにスラッシュ(“/”)を使用した表記も散見される。

[5] 機関の旧名の表記

図表 17 にある Tokyo College of Pharmacy は、東京薬科大学(Tokyo University of Pharmacy and Life Science)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 が、現在名と同程度あるいはそれ以上に使われている。

[6] スペルの違い、その他

純粋なミススペルは除いて、以下のような表記ゆれがある。

- ローマ字書式のゆれ

  九州大学の英語正式名は Kyushu University だが、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。

- 単数形と複数形の混同

  東京薬科大学の正式英語名は Tokyo University of Pharmacy and Life Science であるが、上記リストに見るように、Tokyo University of Pharmacy and Life Sciences 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、東京女子医科大学の Nara Women's University、Tokyo Women's Medical University の“Women's”を“Woman's”と誤記する例も見られる。

- 単語間のハイフン挿入

  横浜市立大学(Yokohama City University)、関西学院大学(Kwansei Gakuin University)、青山学院大学(Aoyama Gakuin University)等では、2 つの単語間にハイフン(“-”)をいれた Yokohama-City University、Kwansei-Gakuin University、Aoyama-Gakuin University という表記も相当数存在する。

- 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 Prefectoral、Prefectual、Prefecture と様々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)。

(3) 大学の下部組織の表記ゆれ

大学の下部組織(学部、大学院研究科、付置研究所、病院、博物館等)の表記は、機関の表記より更にゆれが大きい。ここではこれについて詳しくは述べないが、機関の表記ゆれで述べた[1]～[6]のタイプは、当然下部組織にも存在する。しかし、それ以外に、下部組織の表記に関して特徴的なことが 2 点ある。これらは表記の「ゆれ」とは言えないが、組織の同定を非常に困難にする問題なので、ここで触れておきたい。

一つは、大学の教員が学内の複数の組織に属していることに起因する問題である。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、これは研究院、学系などと呼ばれる(英語では Research Faculty、Institute など)。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。このような並列的組織を設けることは各大学の自由ではあるが、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する場合に大きな障壁となる。

もう一つの問題は、記載されている組織が必ずしも第一層の下部組織ではないことである。つまり、学部名を記載せず学科名が記載されているようなケースである。これも、組織別の集計を難しくする一因になる。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。

4-4 公的機関における機関表記のゆれ

(1) 表記ゆれの程度が大きい公的機関

大学の場合と同様、指標 A と指標 B の値が大きい機関を、それぞれ**図表 18** に示す。

図表 18 指標 A と指標 B の値が大きい機関

(a) 指標Aの値が大きい公的機関

順位	公的機関名	表記数	指標A値
1	(独)科学技術振興機構	86	3.58
2	(独)農業・食品産業技術総合研究機構	18	2.93
3	高エネルギー加速器研究機構	26	2.89
4	(認)日本赤十字社	15	2.88
5	(独)理化学研究所	46	2.82

(b) 指標Bの値が大きい公的機関

順位	公的機関名	表記数	指標B値
1	(独)国際農林水産業研究センター	4	0.81
2	(認)日本赤十字社	15	0.74
3	(独)農業・食品産業技術総合研究機構	18	0.70
4	工業技術院生命工学工業技術研究所	10	0.70
5	(独)農業環境技術研究所	9	0.69

(注) Elsevier Scopus を基に科学技術・学術政策研究所が集計

(注) 現存しない工業技術院生命工学工業技術研究所(現・(独)産業技術総合研究所)があるのは、1996 年以降のデータが分析対象のため。

このうち、独立行政法人科学技術振興機構(JST)の例を**図表 19** に、独立行政法人農業・食品産業技術総合研究機構(NARO)の例を**図表 20** に示す。

このリストから解るように、JST では同機構が行っている研究プロジェクトあるいは研究資金制度の名称(CREST、PRESTO など)、あるいはそれを機関名に含めた表記が多い。

図表 19 独立行政法人科学技術振興機構(JST)の表記ゆれ

機関名	機関ID	セクター	Scopusにおける表記ゆれ	英語正式名	出現度数
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	Japan Science and Technology Agency	○	9749
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	CREST		4208
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	JST		3329
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	PRESTO		2753
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	Japan Science and Technology Agency (JST)		1316
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	Japan Science and Technology Corporation		1151
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	Japan Sci. and Technol. Corporation		1061
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	ERATO		947
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	CREST-JST		470
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	Japan Science and Technology		446
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	Japan Science and Technology Corporation (JST)		424
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	Japan Sci./Technology Corporation		406
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	JST-CREST		354
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	Exploratory Res. for Adv. Technology		181
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	Japan Science and Technology Corp.		142
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	CREST JST		141
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	Japan Science/Technology Corporation		118
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	Japan Sci. and Technol. Corp. (JST)		109
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	SORST-JST		95
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	Core Res. Evolutional Sci. T.		83
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	JST CREST		72
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	Exploratory Research for Advanced Technology		69
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	Core Research for Evolutional Science and Technology (CREST)		64
独立行政法人科学技術振興機構	NID201200542960062	特殊法人・独立行政法人	Japan Science Technology Agency		60

(注)表記バリエーション数が非常に多いので出現度数 60 以上のバリエーションのみ示す。

(注)Elsevier Scopus を基に科学技術・学術政策研究所が集計

一方 NARO では、リストの先頭にある NARO 自身の正式名称は出現度数では 7 番目であり、中央農業総合研究センター(National Agricultural Research Center)をはじめとする同機構に属する研究センターを示す表記が上位を占める。

図表 20　独立行政法人農業・食品産業技術総合研究機構(NARO)の表記ゆれ

機関名	機関ID	セクター	Scopusにおける表記ゆれ	英語正式名	出現度数
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	National Agriculture and Food Research Organization	○	30
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	National Agricultural Research Center		477
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	National Agriculture Research Center		244
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	Natl. Agricultural Research Center		177
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	Natl. Agric. Res. Ctr. for W. Region		90
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	Natl. Agric. Res. Ctr. Tohoku Reg.		83
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	Natl. Agric. Res. Ctr. Hokkaido R.		35
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	National Agricultural Research Center (NARC)		18
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	National Agriculture Research Center for Hokkaido Region		18
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	Natl. Agr. Res. Cent.		15
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	National Agriculture and Food Research Organization (NARO)		14
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	Natl. Agric. Res. Ctr. for W. Reg.		14
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	Natl. Agric. Res. Ctr. Tohoku Regn.		14
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	Tohoku National Agriculture Research Center		13
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	Natl. Agriculture Research Center		12
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	National Agricultural and Food Research Organization		11
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	National Agricultural Research Center for Kyusyu Okinawa Region		11
独立行政法人農業・食品産業技術総合研究機構	NID201200072077485	特殊法人・独立行政法人	National Agriculture Research Center for Tohoku Region		10

(注) Elsevier　Scopus を基に科学技術・学術政策研究所が集計

(2) 公的機関名の表記ゆれのパターン

公的機関においても、大学の項で述べた[1]～[6]のタイプの表記ゆれが多く存在する。しかし、公的機関の表記ゆれが大学のそれよりずっと大きい(指標 A、指標 B のいずれにおいても)主な理由は、このように、下部組織名やプロジェクト名を含んだ表記が多いことによる。

実は、大学においても機関名に下部組織を含めた表記は存在するが、少数であることもあってここでの調査対象からは除いている。しかし、公的機関ではこの種の表記は無視できず、これを除外すると表記の大半が抜け落ちる機関もある。

従って、公的機関の論文をデータベースから調査する場合は、その下部機関の名称も考慮に入れて検索を行う必要がある。このことは大学共同利用機関でも同様で、たとえば、自然科学研究機構に属する分子科学研究所や国立天文台などは、ほとんどそれらの研究所名で表記されている。

4-5 誤同定が起こりやすい表記

機関名に様々の表記ゆれがあると、機関別の集計や分析を行う場合、次の二通りの問題が起こる。

一つは、ある機関に該当するデータを網羅的に得られない、すなわちデータの洩れが生ずることである。しかし、よく出現する表記の辞書があれば、完全とは行かなくても、洩れをかなり低く(2～3%程度に)抑えることは可能である。3－2で述べた機関名辞書や表記ゆれテーブルの公開は、この面での支援になると考えられる。

もう一つの問題はより重大である。それは、ある機関名を、それと名称が似ている機関に誤同定してしまうことである。NISTEP で行っている機関名寄せでもこれはたびたび起こり(全体から見れば低い率ではあるが)、結果を検証しては辞書やアルゴリズムを修正するという過程を繰り返している。

ここでは、誤同定を起こしやすい表記を、いくつかのパターンに分けて示す。

[1] 同一の英語機関名

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。民間企業では、全く無関係の機関が同一の英語名を持つ例が見られるが、大学や公的機関ではほとんどなく、機関名辞書に見られるのは、聖泉大学と清泉女子大学(どちらも Seisen University)のみである。

しかし、統合や改組を行った機関が、日本語機関名は変更したのに英語名はそのままという例はそれ程珍しくない。代表的例では、東京都立大学と首都大学東京はどちらも Tokyo Metropolitan University 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、宇宙科学研究所は独立した国立研究所から独立行政法人宇宙航空研究開発機構の下部組織に移行したが、Institute of Space and Astronautical Science の英語名をそのまま用いている。これらの場合は、論文に表記された機関が旧機関か新機関か厳密に判別するのはほぼ不可能である。論文の発表年からの判断も、移行時期の近傍では曖昧さを伴う。

[2] 下部組織が結合した表記

Scopus や WoS の機関データでは、Kyoto University School of Medicine のように、機関名とその下部組織名が合体表記される場合がある(特に Scopus に多い)。このような表記が、別の機関名と似通っていると誤同定が起こりやすい。特に、前置詞等が省略されていると判別が困難である。

Okayama Univ Sci という表記は、多分岡山理科大学(正式名 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。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なので、Tokyo University of …という英語名を持つ多くの大学と混同しやすい。Tokyo Univ Agr は東京農業大学(正式名 Tokyo University of Agriculture)か東京大学農学部か、Tokyo Univ Pharm は東京薬科大学(正式名 Tokyo University of Pharmacy and Life Science)か東京大学薬学部か、などである。

また、東京理科大学の英語正式名は Tokyo University of Science であるが、旧名の Science University of Tokyo もよく使われるので、Fac Sci Univ Tokyo を、東京大学理学部か東京理科大学かコンピューターで判別するのは難しい。このような例は他にも多数挙げることができる。

[3] ありふれた単語のみから成る機関名

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、誤同定が起こりやすい。[2]と同様、下部組織が結合した表記との混同がされやすい。

分子科学研究所(Institute for Molecular Science)は、大学共同利用機関自然科学研究機構に属する研究所であるが、論文では上部機構の名称は略されることが多い。ところが、Institute for Molecular Science あるいはこれとごく類似した名称の付属研究所を持つ機関は多数存在する(たとえば愛知医科大学の分子医科学研究所(Institute of Molecular Medical Sciences))。そのほかの例として、公益財団法人がん研究会のがん研究所(単に研究所名の Cancer Institute で表示されることが多い)、旧・国立公衆衛生院(The Institute of Public Health)(現・国立保健医療科学院)などが挙げられる。

4-6 機関検索の精度の推定 ― Scopus の所属機関検索機能と NISTEP 表記ゆれテーブルを用いた検索の比較

NISTEP で公開している大学・公的機関名英語表記ゆれテーブル（3－3（2）参照）を利用すれば、データベースからの機関の検索や同定を高い精度で行えると考えられる。

一方、データベース提供側でも機関検索をできるだけ正確かつ容易に行えるような努力を払っていることを、2－3（2）で述べた。たとえば Scopus では、名寄せによって各所属機関に固有の ID（Affiliation ID)を与え、利用者がある機関を指定すると、それに相当する Affiliation ID を含む論文を一括検索する「所属機関検索」の機能を備えている。

この Scopus の所属機関検索（以下「S 検索」という）と NISTEP の機関名表記ゆれテーブルを用いた検索（以下「N 検索」という）の精度を推定するための、簡単な検索実験の結果を以下に示す。

（1）検索実験の方法

検索対象の機関には、機関名表記のバリエーションが多い東京農工大学を選んだ。この大学を著者所属機関に含み、1996～2012 年の間に発表されたすべての論文を Elsevier 社の書誌検索システムである Scopus で検索した。S 検索、N 検索とも同じ日（2014 年 2 月 6 日）に行った。

(a) S 検索

Scopus の「所属機関検索」を選び、“Tokyo University of Agriculture and Technology”（東京農工大学の正式英語名）と入力すると、入力名称に当たる機関名が示され、そこから該当するものを選ぶと相当する文献が回答される。このとき、入力名称から Scopus が認識した Affiliation ID（この場合は 60004853）も表示される。得られた文献集合を、出版年が 1996-2012 年のものに絞り込む。

(b) N 検索

NISTEPの大学・公的機関名英語表記ゆれテーブル（Scopus版）には、東京農工大学に対して、**図表 15** に示す 24 の表記バリエーションが示されている。Scopusの「文献検索」で、検索項目を「著者所属機関の名称」として、**図表 15** に示す機関名表記を一つずつ入力して検索する（この時出版年は 1996-2012 に指定する）。24 の各表記に対する検索結果の論理和検索（OR検索）により求める結果が得られる[1]。

[1] この検索の際、単純に機関名を入力すると、含まれる単語をばらばらにした AND 検索を行うので、正しい結果が得られない。入力した綴り通りの検索を行うには、全体を[ ]で囲む必要がある。

（2）検索の結果

S 検索では 12,776 件、N 検索では 12,506 件の論文が得られた。両方の検索結果に共通に含まれる論文（論理積）は 12,348 件で、ほとんどが重複している。

S 検索のみで得られた論文 428 件と、N 検索のみで得られた論文 158 件をダウンロードし、個々の所属機関データを見て、それぞれの論文が確かに東京農工大学の著者によるもの（正解）か、他の機関が間違って検索されたもの（ノイズ）であるかを判定した。その結果、S 検索の 428 論文中 3 件がノイズであった（2 論文は東京農業大学を、1 論文は東京大学大学院農学研究科を誤認）。N 検索の 158 論文はすべて正解であった。

正解の論文中、S 検索でのみ得られた回答には、東京農工大学を**図表 15** 以外の形で表記したものの他、スペルミス（Agriculture を Agrculture など）や単語間にスペース区切りのない表記も含まれていた。Scopus の名寄せでは、このような表記の誤りも吸収していることが判った。一方、N 検索でのみ得られた回答では、Tokyo Noko University あるいは Tokyo Noko University of Technology の表記が圧倒的に多く、その他に Tokyo Univ. Agric. T.などが見られた。以上の結果を**図表 21** にまとめた。

図表 21　Scopus での東京農工大学の機関検索の結果

		S検索			N検索		
		論文数	Aに対する比	D1に対する比	論文数	Aに対する比	D1に対する比
検索された論文	(A)	12,776	100%		12,506	100%	
S、N両方で検索された論文	(B)	12,348	96.65%	95.49%	12,348	98.74%	95.47%
S、Nのどちらかで検索された論文	(C)	428	―	―	158	―	―
Cのうち正解	(C1)	425	3.33%	3.29%	158	1.26%	1.22%
Cのうちノイズ	(C2)	3	0.02%	―	0	0.00%	―
全検索論文	(D)	12,934	―	―	12,934	―	―
Dのうち正解	(D1)	12,931	―	100%	12,934	―	100%

（注）Elsevier　Scopus を基に科学技術・学術政策研究所が集計

S 検索と N 検索で共通に得られた論文(**図表 21** の B)はすべて正解であり、どちらの検索でも得られなかったものはすべて正解ではないと仮定すれば、S 検索では全正解論文の 98.8%、N 検索では 96.7%が検索されたことになる。また、検索論文中のノイズは、S 検索では 0.02%、N 検索では 0 である。このように、S 検索、N 検索とも極めて高い精度で機関名検索を行えることは、他のいくつかの大学に対する検索例でも確認された。

この結果から、機関の名寄せなどに苦労しなくても、データベース提供機関が用意した機関検索機能を使えば良いではないかと考えられるかもしれない。ある特定の機関の論文だけを集めたいという目的なら、確かにその通りで、Scopus ならば所属機関検索機能を使えば十分であろう。しかし、ミクロデータ分析のように、あるテーマに関する論文を広く検索し、その中の機関別分布を調べたい場合には、ダウンロードした論文集合中の所属機関データを分析する必要があり、この機能は使えない。このような場合には、NISTEP の機関名辞書や表記ゆれテーブルが役に立つと考えられる。

5 まとめ

ここでは、次の 2 点について注記する。一つは、論文執筆者（所属機関や雑誌出版者等の関係者を含む）に対する所属機関表記についての要望である。もう一つは、今後の機関データ整備に関する NISTEP の考え方についてである。

5-1 論文執筆の際の所属機関表記について

論文を発表するときには、著者の所属機関・組織を正確に表記することが求められる。不統一、不正確な表記は、機関や組織の業績評価に不利益をもたらすことになりかねない。

たしかに論文データベース提供機関が、個々の研究機関の名寄せについても積極的に取り組んでいる。しかしながら、大学もしくは部局ごとに今一度英語表記の統一化を図ることで、論文発表に関する意識の向上がなされるのではないだろうか。また、タイムズ社の大学ランキングにおいては論文数のような定量的指標のみではなく、世界の研究者の間での存在感(visibility)に関する定性的な調査結果も含まれている。このような visibility の向上のためにもやはり大学名や部局名を統一化させることが重要ではないだろうか。従って、個々の論文発表者が注意すると同時に、機関全体で統一的表記を定め、構成員にそれを周知徹底することが望ましい。

大学の場合、大学院生や研究員への教育も必要である。現在、多くの大学や研究所では、その構成員の発表論文を機関リポジトリから公開しているので、このような周知・教育には、機関リポジトリの運営に当たっている図書館が関与するのが効率的であるかもしれない。

また、当該の機関の努力だけでなく、論文が発表される学術雑誌においても、正しい統一的表記が受け入れられるように投稿規定を定めることが必要である。

これまでに述べたことの繰り返しになるところもあるが、所属機関表記に当たって特に注意してほしい点をまとめておく。

[1] 機関や組織の正しい名称を正確に表記する。機関名と組織名は明確に分離する（“Faculty Y X University”ではなく、“Faculty of Y, X University”のように）。

[2] 4－3（3）で述べたように、大学の教員が学内の複数の組織に属していることが多いが、論文発表の際はどの組織を記載するか、大学全体で見解を統一することが望ましい。

[3] 著者が 2 つ以上の機関を兼務している場合、研究に外部資金を得ている場合、ある機関から別の機関に派遣されている場合など、著者所属に複数の機関を記載しなければならないことがある。このような場合はそれぞれの機関を分離して記載する。たとえば、X 大学に所属する著者が JST の CREST によって研究を行った場合、“X University, JST CREST”ではなく、“X University”と“CREST, Japan Science and Technology Agency”の 2 つの所属を記載する。

[4] いくつかの大学共同利用機関や独立行政法人では、それらの機構の下にかなり独立性の高い多くの研究所や施設が存在する。このような機構では、機構名と研究所名を併記するか、研究所名のみを記載するかを機構全体で定めた上で、その記法を統一することが望ましい。

5-2 今後の機関データ整備の進め方

NISTEP では、ミクロデータ分析を支える機関データ整備は、大学・研究機関や政府が根拠に基づいて研究開発の方針や計画を策定するための基盤として重要と考え、本報告書で述べたデータ整備活動を今後も継続発展させる考えである。特に留意したい点は以下の通りである。

(1) 機関下部組織のデータの充実

機関名辞書では、研究活動において主要な大学や公的機関の下部組織を登録している(3－2(1)参照)が、論文中の下部組織の表記は極めて多様かつ複雑であり、その表記ゆれの分析や生産統計への実装はいまだ不十分である。しかし、大規模な機関においては下部組織レベルのミクロデータ分析も重要なので、表記ゆれの検討、名寄せアルゴリズムの改善等を図っていきたい。しかし、機関の組織構成は頻繁に変わるので、あまり過去に遡ることは断念し、ある時点以降の組織変遷情報を確実に把握したいと考えている。

(2) インプットデータとアウトプットデータの接続

3－1で概念的に述べたように、NISTEP が進めているデータ整備プロジェクトは、研究インプット(研究者数、研究費等)と研究アウトプット(論文、特許等)のデータを接続させることを主要な目標としている。このため、インプットのデータ源である科技調査名簿と、アウトプットのデータ源である Scopus や WoS を、それぞれ機関名辞書に接続しているが、インプット側とアウトプット側の対応付けはまだ検証していない。両者とも、機関、組織が年とともに変遷するので、それを考慮した対応付けが必要である。これについて整備を進め、両者の接続をサポートするツールを公開したいと考えている。

(3) データベース提供機関及びデータ利用者との交流促進

データベース提供機関でも、機関名や著者名の正確な検索・同定のためいろいろな工夫を行っていることは既に述べたとおりである。これらと連携することにより、NISTEP の整備活動が進むとともに、NISTEP の成果をデータベースサービスに活用する可能性もあると考えられる。

また、NISTEP が整備したデータの利用者や潜在的利用者との交流を進めることにより、利用者のニーズをより的確に把握することが重要である。これらの交流活動は既に試みているが、今後発展させたい。

(4) 継続的データ整備のための方策の検討

ここで述べたデータの整備は、ある期間行えば完了するという性格のものではない。論文等に現れるすべての機関や著者に一意的な識別記号が与えられない限り、継続的な活動が必要である。そのための方策や体制を検討することは、ある意味で最重要な課題である。

参考文献

[1] 富澤宏之,「科学技術イノベーション政策に有用なデータ基盤は何か～世界的動向と歴史的視点からの考察～」, 文部科学省科学技術政策研究所, 科学技術政策研究レビュー, 第 2 巻, pp.46-83, 2012 年 2 月.

[2] 富澤宏之, 岸本晃彦「データ・情報基盤整備に関する課題」, 研究・技術計画学会, 第 27 回年次学術大会, 2012 年 10 月 27 日, (年次学術大会講演要旨集, pp. 102-105).

[3] 文部科学省科学技術政策研究所科学技術基盤調査研究室,「『科学技術イノベーション政策のための科学』におけるデータ・情報基盤構築の推進に関する検討」, NISTEP NOTE(政策のための科学), No.3, 2012 年 11 月.

[4] 富澤 宏之,「論文・特許データの整合的データ体系の構築～データの名寄せの挑戦～」, 文部科学省科学技術政策研究所, 科学技術政策研究レビューセミナー, 第 4 巻, pp.1-26, 2013 年 3 月.

[5] 富澤宏之, 岸本晃彦, 小野寺夏生, 中山保夫, 伊神正貫,「エビデンスベースの政策形成のためのデータ・情報基盤の展開」, 研究・技術計画学会, 第 28 回年次学術大会, 2013 年 11 月 2 日, (年次学術大会講演要旨集, pp. 340-343).

[6] 伊神正貫, 小野寺夏生, 富澤宏之,「大学・公的機関における研究開発に関するデータの整備と公開－SciREX データ・情報基盤構築の成果の紹介－」, 研究・技術計画学会, 第 28 回年次学術大会, 2013 年 11 月 2 日, (年次学術大会講演要旨集, pp. 344-347).

調査体制

本調査の体制は以下の通りである。

（調査実施・報告書執筆）

小野寺夏生　　科学技術・学術基盤調査研究室　客員研究官

（報告書執筆　協力）

阪　彩香　　科学技術・学術基盤調査研究室　主任研究官

（調査実施）

富澤宏之　　科学技術・学術基盤調査研究室　室長

伊神正貫　　科学技術・学術基盤調査研究室　主任研究官

（2014 年　4 月時点）

NISTEP NOTE(政策のための科学) No.11
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－

2014 年 5 月

文部科学省 科学技術・学術政策研究所
科学技術・学術基盤調査研究室

〒100-0013 東京都千代田区霞ヶ関 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