

Title	我が国の大学・公的機関に関する情報のアーカイブ化 ： NISTEP大学・公的機関名辞書の整備
Author(s)	小野寺，夏生；伊神，正貫；富澤，宏之
Citation	年次学術大会講演要旨集，30：1036-1040
Issue Date	2015-10-10
Type	Conference Paper
Text version	publisher
URL	http://hdl.handle.net/10119/13450
Rights	本著作物は研究・技術計画学会の許可のもとに掲載するものです。This material is posted here with permission of the Japan Society for Science Policy and Research Management.
Description	一般講演要旨



我が国の大学・公的機関に関する情報のアーカイブ化 —NISTEP 大学・公的機関名辞書の整備—

小野寺夏生, ○伊神正貫, 富澤宏之(文科省・NISTEP)

1. はじめに

我が国における研究開発の動向を把握し、研究開発に関する方針・政策を国レベル、機関レベルで検討する上で、研究開発を行っている機関に関する情報は、最も基盤的なデータの一つである。研究機関の研究インプットや研究アウトプットのデータを得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別にデータを整理・分析しようとするると種々の困難に直面する。

科学技術・学術政策研究所(NISTEP)では、このような困難を低減し、研究開発に関するマイクロ分析を支援するために、我が国の研究開発機関(大学と公的機関に重点)の基本的情報を収録する「NISTEP 大学・公的機関名辞書」(以下「機関名辞書」という)の作成に取り組んでいる。この事業は、2011 年度から行われている文部科学省の「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」の一環としての「データ・情報基盤の構築」の一部であり、研究開発に関する機関の基本的な情報を継続的に取得し、アーカイブ化することを目標としている。

機関名辞書については第28回の本大会で一部報告した[1]が、本発表ではその後の進捗、特に主要大学の下部組織(部局)情報の整備に重点を置いて、その構成と内容、機関・組織の情報を収集する方法について報告する。また、その活用例として、論文データベース中の所属機関の同定の方法と結果の概要を報告する。

NISTEP では、機関名辞書を各方面で活用していたべく、関連のデータとともに公開している[2]。

2. 機関名辞書の構成と内容

機関名辞書には、研究開発を行っている日本の機関(以下「代表機関」という)とその主な下部組織の名称(日本語と英語)、所属セクター、階層関係、変遷等の情報を収録する。機関名辞書の最新版(2015 年 5 月バージョン)に収録されている機関数を表 1 に示す。

2.1 収録する機関と下部組織

(1) 代表機関

- ① 現存する全ての大学、短期大学、高等専門学校、

大学共同利用機関、及び学校法人

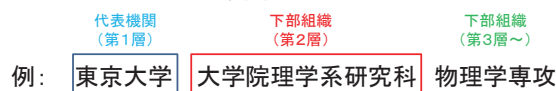
- ② 研究開発を行っている国の機関、特殊法人・独立行政法人(国立研究開発法人を含む)、地方公共団体の機関、非営利法人、その他の機関(日本所在の国際機関等)
- ③ ①または②のセクターに属し、この15年ほどの間に統廃合、名称変更等があって現存しない機関のうち主要なもの
- (上記には民間企業が含まれていないが、これについては2.4を参照)。

(2) 下部組織

上記の代表機関に属する以下の下部組織(現存しないものも含む)を収録する。原則として第2階層下部組織(代表機関の直下の下部組織)までを対象とする¹。図1に代表機関-第2階層下部組織-第3階層下部機関の関係性のイメージを示した。

- ① 研究論文数の特に多い大学の部局(2.3 参照)。
- ② 大学に属する病院、及び共同利用・共同研究拠点となっている部局。これらについては第3階層下部組織(医学部の附属病院のように第2階層下部組織の下組織)であっても収録対象とする。
- ③ 大学共同利用機関である4つの機構に直属する研究所等。
- ④ 国の機関及び特殊法人・独立行政法人等に所属する病院、大学校、及び研究開発活動の分析の観点から重要なその他の下部組織。

図1 代表機関-第2階層下部組織-第3階層下部機関の関係性のイメージ



¹ なお、つぎの機関は下部組織ではなく代表機関としている。a) 大学の一部としての短期大学部、高等専門学校、b) 国立高等専門学校:独立行政法人国立高等専門学校機構の下組織であるが、個々の高専を代表機関としている、c) 国立試験・研究機関:但し、試験・研究機関に属しない国の機関(気象庁高層気象台、厚生労働省の検疫所等)は、属する省庁の下部組織とする、d) 地方公共団体の公設試験研究機関等。

表 1 機関名辞書(2015 年 5 月バージョン)に収録されている機関数

セクター	現存機関			非現存機関		
	代表機関	下部組織	小計	代表機関	下部組織	小計
国立大学	86	1268	1354	15	122	137
国立短大				26		26
国立高専	51		51	8		8
公立大学	85	60	145	16	1	17
公立短大	17		17	45		45
公立高専	3		3	4		4
大学共同利用機関	4	24	28	3		3
国の機関	40	9	49	57	7	64
特殊法人・独立行政法人	88	224	312	37	29	66
地方公共団体の機関	685		685	90		90
学校法人	663		663	1		1
私立大学	612	364	976	41	17	58
私立短大	341		341	223		223
私立高専	3		3			
会社						
非営利団体	3569		3569	317		317
その他	5		5			
計	6252	1949	8201	883	176	1059

2.2 機関名辞書に収録する情報

上記に該当する代表機関及び下部組織に対する以下の情報を収録する。

- ① NISTEP が与える機関 ID (代表機関、下部組織のそれぞれに独立して与える)
- ② 日本語と英語の正式名
- ③ 所属するセクター (17 に分類、表 1 を参照)
- ④ 下部組織とそれが属する代表機関とのリンク
- ⑤ 現存しない代表機関、下部組織の変遷情報 (移行の種別、移行年月日、継承先がある場合その機関)
- ⑥ 郵便番号

2.3 下部組織レベル情報の付加の必要性

研究開発の動向を把握するとき、下部組織レベル (大規模な総合大学の学部・大学院研究科等の単位や大学共同研究機関に属する研究所単位など) での調査・分析が重要な場合がある。2.1(2)で示したように、重要な下部組織も収録することが機関名辞書の一つの特徴である。特に、大学下部組織情報のアーカイブ化を一つの重要な目標とし、当面、1996～2010 年の期間の論文数 (Scopus のデータによる) に基づいて選定した 32 大学の第 2 階層下部組織 (事務組織を除く) の情報を網羅的に収録している。これらの大学と収録されている下部組織数 (2014 年度の現存している下部組織のみ) を表 2 に示す。

表 2 機関名辞書(2015 年 5 月バージョン)に収録されている 32 大学の下部組織数

大学	下部組織数	大学	下部組織数
北海道大学	86	大阪大学	64
東北大学	79	神戸大学	42
筑波大学	67	岡山大学	55
群馬大学	21	広島大学	53
千葉大学	48	徳島大学	39
東京大学	74	九州大学	134
東京医科歯科大学	26	長崎大学	35
東京工業大学	67	熊本大学	38
東京農工大学	24	大阪市立大学	35
新潟大学	27	大阪府立大学	24
富山大学	36	慶應義塾大学	53
金沢大学	50	東海大学	71
信州大学	30	東京理科大学	31
岐阜大学	27	日本大学	51
名古屋大学	67	早稲田大学	90
京都大学	70	近畿大学	56

注: 国立、公立、私立の順番で地域順に示している。

2.4 機関名辞書の補助情報

機関名辞書には、2.1 及び 2.2 に示した収録基準を満たさないが、NISTEP 内部での機関同定や分析の作業上必要性が高い次のような情報を補助情報として含んでいる。これらの情報は、公開される機関名辞書には

含まれない。

a) 2.1 に該当する代表機関及び下部組織の正式名称以外の名称(日本語及び英語の別名、略称、表記ゆれデータ等)

b) 2.1 に該当しない代表機関・下部組織に関する 2.2 に相当する情報

具体的には以下の機関・組織が該当する。

- ・ 民間企業(代表機関、下部組織とも)
- ・ 2.1(1)に示す代表機関の下部組織で、2.1(2)に該当しないもの

民間企業を公開情報に含めていない理由は、別途に構築、公開している NISTEP 企業名辞書に、我が国の企業に関する主要な情報が収録されているためである。

補助情報に含まれる下部組織や正式名称以外の機関名称は、同定漏れや誤同定の原因を防止するために収録されている場合が多く、収録に明確な基準があるわけではないため、公開の機関名辞書には含めていない。しかし、よく使われる別名や略称は「大学・公的機関名英語表記ゆれテーブル」として公開している[2]。

3. 機関名辞書整備のための情報収集

3.1 大学と国に関係する機関

(1) 代表機関の情報

大学、短大、高专、大学共同利用機関、国の機関、特殊法人・独立行政法人、学校法人の各セクターに属する代表機関については、毎年少なくとも 1 回、府省等の Web サイトのリストを調べて存在を確認し、機関名辞書に対し以下の更新処理を行う。

- ① 新しくリストに記載された機関を辞書に登録し、その機関の Web サイトあるいは他の情報源から必要な情報を入手する。
- ② リストからなくなった機関については、変遷の区分(統合、廃止、名称変更の 3 種に分類)、変遷年月日、継承機関(あれば)を調査し、その情報を辞書に追加する。

(2) 下部組織の情報

機関名辞書に含まれる下部組織については、少なくとも年 1 回、それぞれが属する代表機関の Web サイトで存否を確認する。新しい組織、存在しなくなった組織については、代表機関の場合と同様の調査を行う。

2.3 に示した 32 大学の部局情報の収集も基本的にこの方法によるが、第 2 階層組織を網羅的に収録するため特に注意深く行っている。そのため、各大学の部局情報を入手する Web ページ(多くは大学概要のサイトにある組織のページ)を規定し、このページの記載に基づ

き、辞書に登録すべき部局を識別する。これらの Web ページから英語名称や変遷情報を得られない場合は別のページを調査する。

3.2 地方公共団体の機関、民間企業、非営利法人に属する機関

機関名辞書の作成を開始した 2011 年に、これらのセクターに属する機関は主に「全国試験研究機関名鑑」(2008/2009 年版)[3]に基づいて選択した。その後は、主要な非営利団体について財団・社団法人から公益あるいは一般財団・社団法人への移行に関する一斉調査を行った例を除けば、気が付いた時点で適宜、機関の新登録、変遷情報の追加、英語名の追加や変更を行っている程度である。今後は、現状をより正確に把握するため、少なくとも数年に 1 回程度調査とデータ更新を行う必要がある。

ここで示した方法以外に、4 で述べる論文データベースの機関同定によりデータ更新に重要な情報が得られる。同定できなかった所属機関データによく出現する表記から、機関名辞書に未登録の機関や下部組織、未収録の英語別名や機関略称の情報が得られる。

4. 機関名辞書を用いた論文データベース中の所属機関同定

NISTEP では、機関名辞書と論文データベースをリンクさせることによって、我が国の機関の研究活動データの取得と分析を行っている。その際重要なのは、論文データベース中の著者所属機関データと機関名辞書中の機関(代表機関または下部組織)との対応付け(機関同定)である。

ここでは、Web of Science Core Collection (WoSCC) データベースを用いた機関同定について、同定を難しくする問題点と、それぞれの問題点に対処するための機関名辞書や同定アルゴリズムの工夫について述べる。

4.1 機関名の表記ゆれ

【問題点】

論文データベースの所属機関には、同じ機関の名称が様々の形で表記されている(これは、もとの論文における著者所属機関の表記が多様であることの反映である)。そのパターンは次のように大別される:(a) 正式名称以外の通称や略称による表記、(b) 機関の旧名の表記、(c) 単語の省略、脱落、順序の入れ替え、(d) 単語の略記、(e) 冠詞、前置詞、接続詞、記号(ハイフン等)の省略や書き換え、(f) 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、単語間のハイフン挿

入の有無、単純なスペルミスなど)

【対策】

まず、機関名辞書の補助情報(2.4 参照)の充実がある。この対象は主に上記の(a)~(c)であるが(f)にも一部対応している。一方、(d)~(f)に対してはアルゴリズムで対応している(略記辞書の適用、規格化处理、曖昧マッチングなど)。

4.2 代表機関名と下部組織名のサブフィールド配置の不規則性

【問題点】

WoSCC では、所属機関フィールド(機関名の他、所在地、郵便番号、国の情報を含む)から代表機関と下部組織の部分を取り出し、それぞれを ORG、SUBORG というサブフィールドに収録している。これは同定を行う上で便利であるが、(a) ORG に代表機関と下部組織あるいはキャンパス名が結合(例: Univ Tokyo Hosp、Univ Tokyo Hongo); (b) ORG と SUBORG に代表機関名が分離(例: Dokkyo Univ Sch Med の Dokkyo Univ が ORG に、Sch Med が SUBORG に泣き別れ); (c) ORG に下部組織名が、SUBORG に代表機関名が入る、など、この原則から外れた場合がある。

【対策】

同定処理の中核部分である最長マッチング処理では、所属機関データ表記単語列を機関名辞書内表記単語列と比較し、後者を完全に内包する単語列が前者の中に見つかれば、そのうち最も長い単語列のデータに相当する機関名辞書の機関に同定するが、この処理はまず ORG サブフィールドに対して行い、その後、SUBORG と ORG を合体した単語列(SUBORG+ORG)に対して行う。SUBORG+ORG 単語列とのマッチングにより、上記のように ORG と SUBORG へのデータ配置が原則から外れている場合にもある程度対処することができる。

4.3 一つの所属機関フィールドに複数の機関が存在

【問題点】

論文の著者が複数の機関に所属している場合、通常は別々の所属機関フィールドに収録されるが、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。次の例で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ている。

“Div Mammalian Dev, Natl Inst Genetics, Dept Genetics, SOKENDAI”

この著者はその両方の機関に所属していると考えら

れるので、両方の機関を同定できることが必要である。

【対策】

最長マッチングでは、SUBORG+ORG 単語列で一つの機関がマッチした後、単語列の残りの部分が別の機関にマッチしないかサーチし、全単語列をサーチするまでこの手続きを繰り返す。

4.4 同一の英語機関名を持つ機関

【問題点】

統合や改組により日本語機関名は変更したのに英語名は旧機関と継承機関で同じという例がよく見られる(たとえば東京都立大学と首都大学東京はどちらも Tokyo Metropolitan University)。このような場合、新旧どちらの機関に対応するか判断しづらい。

無関係の機関が同一の英語名を持つ例は会社に多少見られるが、大学や公的機関ではほとんどない。

【対策】

同名の機関が存在して両者の間に継承関係がある場合は、機関名辞書の移行年と論文の発行年を比較し、発行年が移行年以前なら旧機関に、そうでなければ新機関に同定する(移行年データがない場合は新機関に同定する)。但し、統合により代表機関が継承機関の下部組織になる等いろいろな場合があるので、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た。

関係のない機関の英語名がたまたま同じ場合は、郵便番号比較により識別するが、そのデータがなく決められない場合は同定不能とする。

4.5 英語名が非常に似ている機関

【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、類似した名前の機関が存在して誤同定が起りやすい。たとえば、分子科学研究所(Institute for Molecular Science)に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。

似通った名称の大学の間でデータベース中の表記が誤っていたり曖昧であったりすると、違う大学に同定されてしまう。4.2 に示した Dokkyo Univ Sch Med はその例で、獨協医科大学でなく獨協大学に誤同定される可能性がある。

【対策】

ありふれた単語からなる機関の誤同定防止のためには、類似の名称の機関(下部組織を含む)に使われるいろいろな表記を機関名辞書に補助情報として収録する。

また、似通った名称の大学については次のような特別措置を適用している。それぞれの大学に正しく同定さ

れた機関データ(その判定には郵便番号を利用)に含まれる単語列から、その大学との共起性が高いものを抽出した辞書を作成し、この辞書を同定に活用する。たとえば静岡大学と静岡県立大学の場合、“Fac Sci”や“Fac Agr”は前者、“Sch Pharm Sci”や“Sch Food Nutr Sci”は後者に関係づけられる。

4.6 大学の第2階層下部組織同定における問題

【問題点】

32大学の第2階層下部組織の同定においても、4.1で述べた表記ゆれの問題がある(むしろ更に著しい)が、それ以外に第2階層下部組織同定を難しくする独自の問題がある。それは、大学直下の第2階層下部組織を省略してその下の組織が記載されることがしばしばあることである。たとえば、“Graduate School of Science, X University”ではなく“Department of Biochemistry, X University”のような所属表記がなされる。

【対策】

上記のような場合に対する部局同定率の向上のために、以下の改善策を実施した²。

(1) 第2階層下部組織統計辞書の開発

正しい第2階層下部組織同定がなされた機関データ(第2階層下部組織が記載されている)から、大学別に出現頻度の高い単語列を抽出し、これらの単語列に対して、第2階層下部組織別の共起率(大学全体の出現頻度のうちその第2階層下部組織との共起の割合)を求める。出現頻度が高く、かつある特定の第2階層下部組織に共起が集中している単語列を、その第2階層下部組織に関係づけたものが第2階層下部組織統計辞書である。

こうして作成した第2階層下部組織統計辞書を、最長マッチングで32大学に同定されたがその第2階層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第2階層下部組織名とはマッチしなかったデータ)に適用する。たとえば、東京大学に同定されたデータに“Ctr Human Genome”や“Mol Med Lab”が含まれれば医科学研究所、“Dept Ophthalmol”や“Dept Surg”が含まれれば医学系研究科である確率が高いとして、それらの第2階層下部組織に同定される。

(2) 補助情報の対象拡張

第2階層下部組織統計辞書の対象となった第2階層下部組織表記のうち、特定の第2階層下部組織を示すことが間違いないと人手で判断できる表記は、機関名辞

書の補助情報に加えた。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。

5. おわりに

機関名辞書の今後の重要課題として、その継続的維持がある。機関やその下部組織は常に変遷しており、その情報を確実にとらえてアーカイブ化するためには、不断の状況把握・情報更新が必要である。セクターごとにその効率的方法を定め、そのための体制と予算措置を継続的に講じられるよう検討を進めたい。

もう一つの重要課題として大学部局情報の拡張がある。現在部局情報の網羅的収録の対象としている32大学は、1996～2010年の期間の論文数を基準として選定しているが、その実態は年とともに変化するので適宜見直しが必要である。本来は、ある程度の研究規模を持つ総合大学をすべて対象とするのが望ましい。しかし、現在の32大学だけでも、毎年各大学の第2階層下部組織の変遷を調査することはかなりの負担であり、その作業を軽減するための方法を今後検討したい。

謝辞

ここで述べたデータ整備事業は、指定した作業を請け負った株式会社RNAiの協力の下に行われた。

参考文献

- 1) 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).
- 2) “データ・情報基盤—政策研究の高度化とエビデンススペースの政策形成のためのツール.” 科学技術・学術政策研究所. <http://www.nistep.go.jp/research/scisip/data-and-information-infrastructure>.
- 3) 全国試験研究機関名鑑 2008-2009. 文部科学省科学技術・学術政策局監修, 丸善, 2008.

² これらの改善策を実施しても、東京医科歯科大学については、第2階層下部組織同定が困難である。したがって、現状で、第2階層下部組織についての分析が可能なのは、31大学となっている。