１Ｉ１１

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－

○伊神正貫, 小野寺夏生, 富澤宏之（文科省・ＮＩＳＴＥＰ）

1. はじめに

科学技術・学術政策研究所(NISTEP)では、2011 年度から、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」(SciREX)の一環として「大学・公的機関における研究開発に関するデータ整備」を進めている。

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[1]の研究開発に関するデータの整備を行うことを目的としている。

特に、研究開発インプットとアウトプットのデータをミクロレベルでリンクさせ、政府の研究開発投資の成果や研究開発システムに与える影響を、定量的・構造的に分析できるようにすることを目指している。本報告では、この事業によるデータの整備と公開の状況、整備したデータの活用事例、及び今後の予定について述べる。

2. データ整備の現状

2.1 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「機関名辞書」という）は、研究インプットデータとアウトプットデータのミクロ接続のための中核的役割を果たす。

(1) 収録対象の機関

研究開発を行っている国内の機関が対象である。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、全部で約１万２千機関（後述する下部組織や非現存機関を含む）に達している。それぞれの機関には、NISTEP 独自の識別 ID を与えている。

表１ セクターと属する機関数

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

注：機関数には統廃合、改組、名称変更等で現存しない機関も含んでいる。

(2) 収録する情報

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している。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。

② セクター

各機関は、表１に示す 16 のセクターに分類されている。表 1 には、現在公開している機関名辞書に含まれる機関数も示す（現存しない機関を含み、下部組織を含まない）。また、これとは別に、病院にはそのことを示すフラグを与えている。

1 この事業でいう大学には短大、高専、大学共同利用機関を含み、公的機関には国の機関、特殊法人・独立行政法人を含む（地方公共団体の機関は含まない）。

③ 機関の下部組織

一部の大学については学部・研究科・付置研究所等、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している。上部機関と下部組織の間に関係づけをしている。

④ 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、できるだけ収録している。これらについては、変更のあった日付、継承機関(存在する場合)等の情報も収録した。

⑤ 所在地の情報

機関同定のための補助情報として、郵便番号等の情報を収録している。

2.2 研究アウトプットデータベースにおける機関同定:機関名辞書とのミクロ接続

研究論文、特許、学会発表等、研究アウトプットのデータベースは数多いが、現在までに、世界的な論文データベースであるScopus及びWeb of Science (WoS)と、機関名辞書とのミクロ接続を行った。ここでは、論文データベースとの接続の詳細について述べる。

(1) 対象としたデータ

Scopus と 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の同定(機関名辞書の登録機関への対応づけ)を行った。WoSについては、日本以外の5か国(米、英、仏、独、中)の主要大学(各国約200)の論文についても機関同定を行った。これは、日本の大学と論文生産の比較を行うためである。

(2) 機関同定の問題点

ScopusもWoSも、同一機関の表記は必ずしも統一されていないため、機関同定アルゴリズムは非常に複雑になる。次のような場合がある。

① 正式の機関名の他に種々の通称や略称が使われる。また、正しくない名称が使われていることも多い(University of Tokyoが正しいがTokyo Universityが使われるなど)。

② 機関とその部局が一体化して表記される(Kyoto University School of Medicineなど)。

③ Scopus では、アドレスデータの中の所在地、機関名、下部組織名の順序が一定しておらず、機関名の識別が難しい。WoS は順序が一応統一されているが、それに従っていない場合もある。

④ 1つのアドレスデータレコードに2つの機関名が入っていることがある。主に一人の著者が 2 つの機関に所属している場合に起こるが、異なる著者の所属機関が混在している場合もある。

⑤ 機関の英語名が変更されることがある。

これらの表記ゆれの他に、類似の名称の機関の存在が同定を困難にすることがある。静岡大学(Shizuoka University)と静岡県立大学(University of Shizuoka)、浜松大学(Hamamatsu University)と浜松医科大学(Hamamatsu University School of Medicine)などがその例である。また、Fac Sci Univ Tokyo を、コンピュータで東京大学理学部か東京理科大学か判別するのはかなり困難である。

このような事例を分析して試行錯誤的に対処することにより、最長マッチングと類似性計算に基づき、NISTEP大学・公的機関名辞書を用いて機関同定を行うアルゴリズムを開発した。

(3) 同定の結果

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、Scopus では 91.9%、WoS では93.6%が機関同定できた。機関同定はできなかったがセクターが確定できたのは、それぞれ1.6%、1.3%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。

2.3 論文生産統計のためのテーブル設計

2.2 で述べたミクロ接続の結果を集計することにより、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。

① 機関別集計とセクター別集計: 機関別の場合、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。

② 年別、分野別、及びこの両者を組み合わせた集計

③ 複数機関共著論文に対し、完全計数法による集計と分数計数法による集計。

2.4 論文謝辞からの研究資金源の予備的分析

WoS では、論文の謝辞(Acknowledgments)に現れる研究資金受給情報を、2008 年後半から収録している。まだ予備的段階であるが、このデータを用いて、日本の研究資金提供機関・制度について集計・分析を行った。

3. データ公開の現状

この事業で整備したデータはできる限り公開し、関係者に利用していただく考えである。他の事業によるデータと合わせ、NISTEP の Web サイトの「データ・情報基盤」のページ<http://www.nistep.go.jp/research/scisip/data-and-information-infrastructure>において、公開を進めている。

2013 年 9 月現在、次の 3 種のデータを、内容と利用法を説明したマニュアルや活用事例とともに公開している。

(1) NISTEP 大学・公的機関名辞書(ver.2012.1)

2.1 で述べた機関名辞書の公開版である。研究活動を行っている我が国の機関(約1万2千機関)を掲載している。この辞書は、個別機関レベルの分析のための基礎情報源として使用することができる。

(2) 大学・公的機関名英語表記ゆれテーブ(ver.2013.1)

2.2 で述べたように、Scopus にしても WoS にしてもその中の著者所属機関には様々な表記ゆれがある。現在、Scopus 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、提供元であるエルゼビア・ジャパン株式会社の了解を得て公開している。このテーブルに含まれるのは、Scopus 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 回以上出現した表記バリエーションである。

表 2 Scopus における筑波大学の表記バリエーションの例

機関名	表記バリエーション	英語正式名	出現度数
筑波大学	University of Tsukuba	○	31,794
筑波大学	Tsukuba University		822
筑波大学	Univ of Tsukuba		327
筑波大学	Univ. of Tsukuba		282
筑波大学	The University of Tsukuba		57
筑波大学	Univ. Tsukuba		18
筑波大学	Tsukuba Univ.		12
筑波大学	Tsukuba Univ		11

表 2 に筑波大学の表記バリエーションの例を示す。多くは英語の正式名称である「University of Tsukuba」と表記されているが、それ以外も約 1,500 論文存在する。これは正式名称の論文数の約 5%に対応しており、大学からの研究開発成果を把握する上では小さい量ではない。

大学・公的機関名英語表記ゆれテーブルは、約 250 の大学・公的研究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、Scopus で機関名検索を行う場合の有効な補助ツールになる。

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

2.2 で述べた機関同定の結果、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。エルゼビア・ジャパン株式会社の了解を得て、Scopus について、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。

① 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
② 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)
③ 機関別又はセクター別の論文生産統計の作成と分析

4. データの分析事例

機関名辞書と WoS の接続による論文生産に関するデータ分析の例をいくつか示す。いずれも、論文数の算出には分数計数法を用いている。

(1) 論文生産の機関集中度の変化

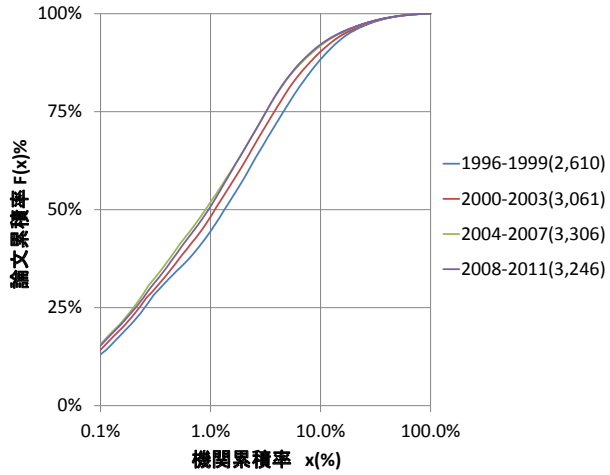
論文数の順に機関を並べ、上位 x%の機関による累積論文占有率 F(x)を求めることにより、機関集中度を見ることができる。1996-2011 年を 4 年ずつの 4 つの時期に分け、時期ごとの x-F(x)関係を図 1 に示す。どの時期

においても、上位 1%の機関で全論文の約 50%、上位 10%の機関で約 90%を占めるという傾向は変わっていない。

(2) 各セクターの生産論文数の変化

各セクターの論文数の変化の状況を図 2 に示す。ここで、国立大学、公立大学、私立大学にはそれぞれの短大、高専を含む。国立大学の占有率はどの時期も約 50%でほとんど変わらない。国立機関の独立行政法人への移行、民間企業の論文生産の低下の傾向が見てとれる。

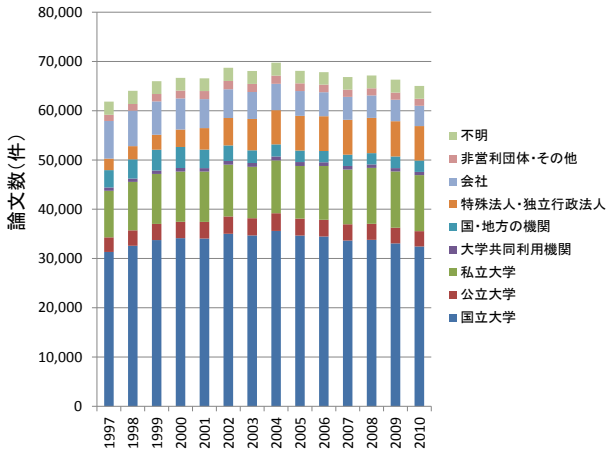
図 1 論文生産の機関集中度の変化



注: 分析対象は、Article, Article & Proceedings (Article として扱うため), Letter, Note, Review のなかで、著者所属に Japan を含むレコードである。凡例のカッコ内の数は、各期間の機関数を示している。

資料:トムソン・ロイター社 Web of Science (SCIE, CPCI: Science)[2011 年 12 月末抽出]を基に、科学技術・学術政策研究所が集計。

図 2 セクター別論文数の変化



注: 分析対象は、Article, Article & Proceedings (Article として扱うため), Letter, Note, Review のなかで、著者所属に Japan を含むレコードである。3 年移動平均の値を示している。

資料:トムソン・ロイター社 Web of Science (SCIE, CPCI: Science)[2011 年 12 月末抽出]を基に、科学技術・学術政策研究所が集計。

5. おわりに ー今後の整備と公開の計画

2012 年度までに 2. で述べたデータ整備を行ったが、その経験の中で、データの網羅性や精度の向上について様々な示唆を得た。そこで 2013 年度は、過去 2 年間の成果を下記の視点から評価し、それに沿ってデータと手法の改善を図ることとしている。

① 機関同定及び異種データ源間のミクロ接続の精度の向上策
② 機関同定等のアルゴリズムの不備と改善可能策
③ 機関名辞書のデータ整備のために必要な外部情報源
④ これまでに生じた各種のエラーへの対処策
⑤ 論文生産等の統計において、必要性の高い集計・分析が効果的に行えるデータ設計
⑥ データ利用者のための補助情報(利用マニュアルやテーブル定義情報)の整備
⑦ データ整備を将来継続するための作業のマニュアル化

データ公開については、3. で述べたように現時点では主に 2011 年度に Scopus から抽出したデータを対象にしている。今年度中には、2012 年度に行った WoS データに関するものも公開する計画である。また、機関名辞書についても、2012 年度に行った新たな機関登録、下部組織や機関変遷情報の拡充等を反映した更新版を公開する。その後は、上述の整備計画の進捗に合わせてできる限り公開を進めていきたい。

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

科学技術・学術政策研究所, データ・情報基盤ー政策研究の高度化とエビデンスベースの政策形成のためのツール, http://www.nistep.go.jp/research/scisip/data-and-information-infrastructure.