特許データベースの開発とイノベーション研究(注1)

後藤　晃（Goto Akira）
東京大学 先端科学技術研究センター　教授

元橋 一之（Motohashi Kazuyuki）
東京大学 先端科学技術研究センター　助教授

1 特許データベースの意義

技術知識を生み出し、学習していくこと、そのような能力を獲得し、維持し、強化していくことは、人々の生活水準、生活の質を維持し、向上していくために決定的な重要性をもっている。個人にとっても、企業にとっても、国のレベルでも同様である。しかし、技術知識について検討するにあたって、その基礎となるデータのソースはきわめて限られている。これは、知識を計測する、ということに本質的に伴う困難さに由来するものである。そのなかで、特許データは技術知識に関する数少ない貴重な体系的な情報の源である。特許制度は、発明へのインセンティブを確保するために一時的に独占的な権利を認めるものであるが、そのためには、発明した後にできるだけ早く出願しなければならず、出願後は18ヶ月後に公開される。結果的に特許は膨大な貴重な技術情報の集積となっている。

特許を発明にかかわるデータとして体系的に利用して本格的な研究を行った古典として、シュムックラーの「発明と経済成長」（注2）があげられる。このなかで、シュムックラーは主として資本財にかかわる個別的な分野の特許の数を時系列的にあらわすと、多くの分野である一時点で特許の数が急増する山が見られることを明らかにした。さらに、その上で、その発明の活発な時期は、その技術分野にかかわる財に対する需要が大きく伸びている時期に対応していることを見出した。企業の研究開発は需要が伸びている分野で活発に行われるので、イノベーションは需要に引っ張られ

る形で起こることになる。このような見方は、科学的な発見がやがて産業に応用されることによってイノベーションが起こる、という見方と対立するもので、シュムックラーの見方はイノベーションは単に科学的知識が実用化されてくるというのではなく、きわめて経済的な活動の一環であることを明らかにしている。

シュムックラーの先駆的研究は1966年に発表されイノベーションプロセスの研究に大きな影響を与えたが、同時に、イノベーション研究の基礎データとしての特許の重要性を広く気づかせることとなった。しかし、特許データは特許の数が多いこと、技術的な内容や、制度の理解など利用にはさまざまな知識が必要なこと、などの理由で、広範には利用されるにいたってはいなかった。

しかし、近年、特許データを利用した経済学的な研究は飛躍的な発展を遂げつつある。経済学のジャーナルには特許データを利用した研究論文が多く載っており、また、特許データを利用したイノベーション研究についてのシンポジウムなども活発に開催されている。

近年、特許データおよびそれを用いたイノベーション研究が活発化した背景としては以下の3点があげられる。第一に、技術革新に対する関心の高まりが背景としてあげられる。技術が国や企業の競争力の重要な要因のひとつである、という認識が広まり、企業経営者、政策立案者などで技術革新に対する関心が高まった。また経済学においても、マクロ経済学における内生的成長論や、応用ミクロ経済学の分野における技術革新の経済分析への関心が高まった。第二に、その技術革新を分析するための重要なデータである特許データがデータベース化され、利用しやすくなったことである。各国、地域の特許データが電子化され、利用しやすくなってきた。後述するように、主として個別の特許の検索を目的としているため、統計データとして利用するためにはさらに編集する必要があるものの、手間とコストをかければ統計的な分析ができるようになってきている。第三に、上の点と密接にかかわっているが、コンピュータ、ソフトウェア、インターネットの進歩によって大量のデータの利用が容易になり、また質的な情報を統計的に処理することも可能になった。

これらの要因が重なって、近年、特許データを利用したイノベーション研究が急速に進展している。現在のところ、米国において特にNBER(注3)の研究者を中心として活発な研究が進んでいる。このことの背景には多くの優秀な研究者が存在していることとともに、NBERの特許引用データベースの存在がある。このデータベースが利用可能になることによって、特許データを用いたイノベーション研究は急速に進んだ。日本や欧州の研究者もこのデータベースを利用するしかない状況であった。日本の特許のうち、米国に出願されるのは一部であり、重要なもの、あるいは外国市場にかかわるものが多いといったバイアスがあるものと思われる。米国特許をもとにしたNBERのデータベースを用いることにはそれ固有の利益もあるが、日本の発明の全体像を明らかにしていくには日本の特許を用いるほうが望ましい。欧州においても研究者によってEPO(注4)のデータを用いた分析が進展している。更にOECD(注5)においては日米欧三極における特許データを統合したパテントファミリーデータが作成され、特許データベースの整備は国際的にも急速に進んできている。

特許データはノイズも多いが技術革新について研究する際には必須の豊かな情報を含んだ基礎的なデータである。日本の特許を用いたデータベースを構築することは、技術革新の研究の進展に、さらには特許制度や技術政策についてエビデンスに基づいた質の高い論議をするためにきわめて重要な意義をもっている。

2 特許データベースの構築

以上を背景に、東京大学先端科学技術研究センターの後藤研究室では、多くのかたがたの協力を得ながら、日本の特許をもとにした特許データベースの構築を行っている。これを(財)知的財産研究所のウェブサイトにおいて近日中に研究用に公開する予定である(注6)。このデータベースが広く利用され、日本においてエビデンスにもとづいたイノベーション研究や特許制度、技術政策の議論が進化し高度化していくことが期待される。以下ではその作成の過程およびデータベースの概要を述べてみたい。

われわれのデータベースの基礎としたのは特許庁の整理標準化データである。われわれはまず整理標準化データの特許についての書誌情報を中心にデータの整理を行い、ついで、それをベースにイノベーション研究に必要な項目をまとめて編集したデータベースであるIIP特許データベースを構築した。

整理標準化データは、(独)工業所有権情報・研修館によって、特許庁で生成される審査経過情報等の各種情報をSGML形式等のデータに整理標準

化して、提供されているものである。これには約902万件の特許が含まれている。整理標準化データでは出願、審査、登録などの個別イベントが発生するごとに登録されているので、延べ件数では約5,600万件のデータを取り扱うことになる(注7)。

このオリジナルの情報は膨大なテキストデータであり、そのままでは計量分析に用いることが困難である。その分量は数百Gバイトになり、パソコンによる統計処理を行うためのデータサイズとしては適当ではない。そこで、まず共通的に利用される変数を抽出した。

しかしながら、この段階でも依然として数Gバイトの容量となり、かつSQLリレーショナルデータベースの形態となっているので、一般的な研究者にとって使い勝手のいいものとはいえない。また、整理標準化データは、更新時点における状態のデータが蓄積されたものであることから、出願人や権利人コードが更新時点によって異なるという問題がある。更に、整理標準化データという特許庁における内部的なデータがベースとなっているものなので、審査請求などの特許に関する処理情報が中間コードという形態で保存されており、分析に用いるためにはデータの変換手続きが必要となる。そこで上記のような点について処理を行い、編集してより使いやすいものにしたものが近く公開を計画しているIIP特許データベースである。特許に関する公開データとしては米国特許をベースとしたNBER patent databaseが有名であるが、IIP特許データベースは日本の特許を用いた、このNBER patent databaseに対応するデータベースである。

3 IIP特許データベースの概要とその構築方法

IIP特許データベースは1964年1月から2004年1月までに出願公開・登録された特許について後述するデータを抽出し、統計処理を行いやすくするためのデータ変換を行い、データベース化したものである。データベースは特許出願ファイル、特許登録ファイル、出願人ファイル、権利者ファイル、引用情報ファイルの5つに分かれている。このそれぞれに含まれている項目は以下のとおりである。

(1) 特許出願ファイル

①出願番号
②出願日
③審査請求日
④出願人番号
⑤請求項数（出願時）
⑥公開・公表の筆頭IPC
⑦公開・公表の筆頭IPCに基づく統合技術分類

(2) 特許登録ファイル

①出願番号
②登録番号
③登録日
④権利消滅日
⑤権利者番号
⑥請求項数（査定・登録時）
⑦公告の筆頭IPC
⑧公告の筆頭IPCに基づく統合技術分類

(3) 出願人ファイル

①出願人番号
②出願人名称
③個法官コード
④国コード及び県コード
⑤出願人登録制度による出願人コード

(4) 権利者ファイル

①権利者番号
②権利者名称

(5) 引用情報ファイル

①引用特許出願番号
②被引用特許出願番号
③引用タイプ（審査官引用又は特許公報引用）

なお、これらのファイルは最大でも数百メガバイト程度であり、インターネットでダウンロードして通常のパソコンで処理できる容量に抑えることができた。

4 IIP特許データベース作成のプロセスとその内容

以下ではIIP特許データベースに含まれている個々のデータについて説明するとともに、整理標準化データからIIP特許データベースに変換したプロセスを記述する。IIPデータベースを利用する際には、必ず、以下のような取り扱いをしていることを理解したうえで利用することが必

要である。

ここでは、（1）特許出願・登録ファイル、（2）出願人ファイル及び（3）引用情報ファイルにわけて述べる。

（1）特許出願・登録ファイル

整理標準化データには1963年以前のデータも収録されてはいるものの、それらについては、ごく一部のデータしか存在しないため、特許登録や出願人などの各種情報が利用可能になる1964年1月以降に出願されたデータを用いることとした。なお、現在、カバーされているデータとしては、2004年1月に出願公開されたものまでが含まれている。(注8) 特許出願・登録データに関して行った主な処理と留意点については以下のとおりである。

①技術分類

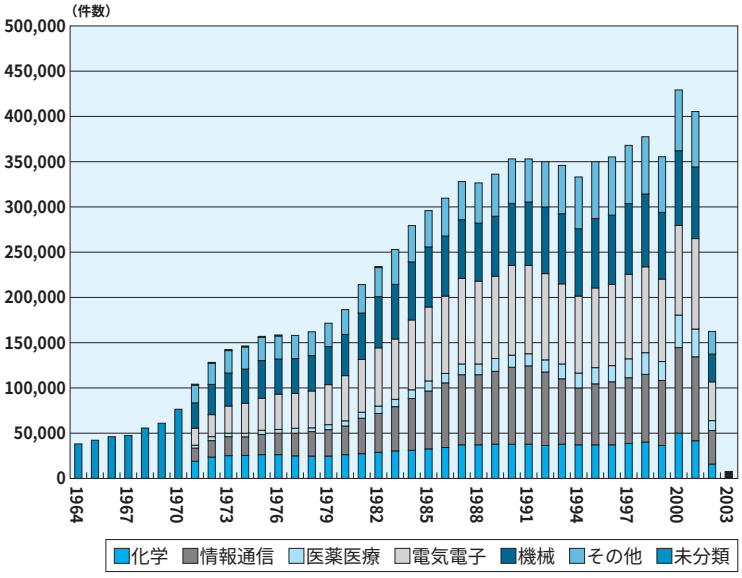
IIP特許データベースにおいては、主分類と副分類が区別されている出願時（あるいは登録時）の筆頭IPCコードを用いて特許の技術分類としている。なお、IPC分類については技術分類の改訂が行われてきているが、IIP特許データベースでは改訂による分類の見直しは行っていない。サブクラスにおりた詳細な技術分類を用いた時系列分析を行う際にはこの点に留意することが必要である。

IPCコードは非常に詳細な技術分野を示すものとして有益であるが、例えばマクロで見た技術分野別出願状況などの全体的な状況を見るためには細かすぎるという問題がある。そこで、公開用データベースにおいては、WIPOの公式統計で用いられている統合技術分類に準じた、表1の技術分類も用意した。また、この分類を更に統合してNBER patent databaseにおける大分類（①Chemical、②Computer and Communications、③Drugs and Medical、④Electrical and Electronics、⑤Mechanical、⑥Others の6分類）と整合的な技術分類も加えて、日米比較が容易に行えるようにしている。また、図1は大分類で見た出願件数を出願年別にグラフにしたものである。

●表1：WIPO統計をベースとした特許技術分類

技術分類	米国対応大分類	内　容	対応IPC
1	6	農水産	A01　（但しA01Nを除く）
2	6	食料品	A21～A24
3	6	個人・家庭用品	A41～A47
4	3	医療機器・娯楽	A61～A63　（但しA61Kを除く）
5	3	医薬品	A61K
6	1	処理、分離、混合	B01～B09
7	5	金属加工、工作機械	B21～B23
8	5	切断、材料加工、積層体	B24～B32　（但しB31を除く）
9	6	印刷、筆記具、装飾	B41～B44
10	5	車両、鉄道、船舶、飛行機	B60～B64
11	5	包装、容器、貯蔵、重機	B65～B68
12	1	無機化学、肥料	C01～C05
13	1	有機化学、農薬	C07、A01N
14	1	高分子	C08
15	1	洗剤、応用組成物、染料、石油化学	C09～C11
16	3	バイオ、ビール、酒類、糖工業	C12～C14
17	3	遺伝子工学	C12N15/
18	5	冶金、金属処理、電気化学	C21～C30
19	6	繊維、繊維処理、洗濯	D01～D07
20	6	紙	D21、B31
21	6	土木、建設、建築、住宅	E01～E06
22	6	鉱業、地中削孔	E21
23	5	エンジン・ポンプ・工学一般	F01～F04、F15
24	5	機械要素	F16～F17
25	6	照明、加熱	F21～F28
26	6	武器、火薬	F41～F42、C06
27	4	測定・光学・写真・複写機	G01～G03
28	2	時計・制御・計算機	G04～G08
29	2	表示・音響・情報記録	G09～G12
30	4	原子核工学	G21
31	4	電気・電子部品、半導体、印刷回路、発電	H01～H02、H05
32	2	電子回路・通信技術	H03～H04
33	6	その他	B81、B82

●図1：技術分野別出願数の推移



なお、整理標準化データにおいてはこれらの公開された技術分類情報のほか、特許庁内において特許検索を行うときに用いられる検索用ＩＰＣコードに関する情報も存在する。このデータは技術分類が改訂されると過去のデータも最新の分類に変更されるため時系列分析を行う際には有効である。しかしながら、特許一件に対して複数存在する検索ＩＰＣコードについて筆頭ＩＰＣという概念のものが存在しないので、全体で約5，６００万件存在する膨大な技術情報をどう処理するかという問題が残っている。この点については、今後のデータベース改良作業のなかで取り組んでいく予定である。

②審査請求日

審査請求制度は１９７１年から導入されたが、審査請求に関する情報は整理標準化データの中では出願特許に関する各種手続き（特許の補正、優先権証明請求など）に関する中間コードとして取り扱われている。審査請求日に関するデータは、審査請求を示す中間コード（Ａ６２１：出願人による審査請求、及びＡ６２５：他人による審査請求）のイベントが発生した日付から作成した。ただし、特許によってはこのイベントが複数回存在するものがあるが、その際には最初に発生した日付を用いた。

③請求項数

請求項数については、出願人による出願時のものと特許庁における審査後の登録時のものは違う場合があるので、それぞれを出願ファイルと登録ファイルに収録している。なお、これらのデータについては、１９７０年以前はほとんどの特許について情報が欠落しており、逆にほとんどの特許について利用可能となるのは１９７５年以降であることに留意が必要である。ただし、改善多項制が導入され、ひとつの特許のなかの項数が全般的に増加し始めるのは１９８７年以降であるので、この点は実質的にはそれほど大きな問題ではないといえよう。

④出願人番号及び権利者番号

個々の特許における出願人及び権利者の情報については、出願人及び権利者のそれぞれに番号を付与し、名称、居住地などの詳細情報については、別のファイル（出願人ファイル及び権利者ファイル）を設けている。１つの特許の複数の出願人又は権利者が存在する場合は、複数の番号を；（セミコロン）で区切って掲載している。

（２）出願人ファイル

出願人ファイルにおいては名称の他、出願人のタイプ（個人、法人又は官庁：個法官コード）や居住地の属性（国コード・県コード）に関する情報が含まれている。個法官コードと国・県コードは特許庁の出願人登録制度に基づく出願コードに従って管理されていることから、出願人コードについて以下の問題に対処する必要がある。

・出願人コードは時系列的に改訂が何度か行われており、過去のデータについては最新時点のコードに統一する必要がある。

・特許庁における出願人コードによる出願人の管理は、１９９２年に申請者登録制度が始まって本格的に始まったもので、それ以前については、数百社の大手企業以外は出願人番号が割り当てられていない。

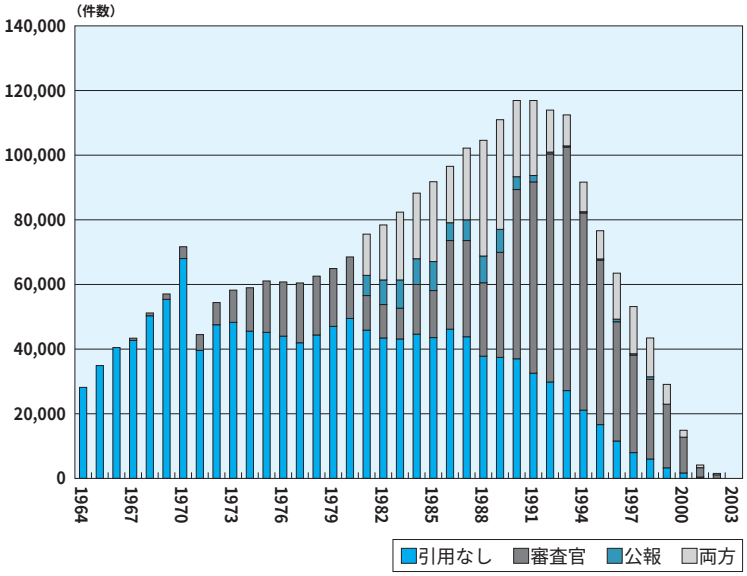
最初の問題については、出願人番号の改訂の履歴を追うことによって過去のデータについても最新の出願人番号（９桁コード）に変換することが可能である。２番目の問題は出願人番号が付与されていない出願人に名称を用いて名寄せを行い、出願人番号を振りなおす作業を行った。更に、特許庁における出願人番号管理の方法として、同一企業であっても異なる事業所から出願が行われる等の場合、別企業として認識されてしまい、異なる番号を付与するケースが存在する。このような場合は企業単位で１つの番号が付与されるよう名寄せ作業を行った。ただし、これらの名寄せ作業は大手の出願人については目視で行ったものの、その他については出願人名称を用いた機械的作業で行ったため、データとしては完全なものではないことに留意が必要である（注９）。

（３）引用情報ファイル

引用情報は、技術の発展のプロセスをたどったり、特許の重要度を推し量ったりするなどさまざまな利用可能性をもつ特許情報のなかでももっとも重要な情報のひとつである。しかし、日本の特許制度のもとでは、その有効性は限られたものとなっており、また日米の制度の違いから、広く用いられている米国特許における引用文献情報とは性格が異なっていることに留意する必要がある。これを用いる場合にはとりわけ以下の点を理解したうえで利用することが必要である（注10）。

引用文献は特許文献と非特許文献に分類されるが、整理標準化データに収録されているのはそのほとんどが特許文献である。なお、例外的に実用新案が先行文献として引用されているものはかな

●図2：登録特許における引用特許数の推移



りの件数に上るが、ここでは国内特許文献に関する引用情報のみに限定してデータを作成している。

整理標準化データには国内特許文献に関する引用情報として以下の2種類のデータが存在する。第一は、審査官が拒絶理由として出願人に提示した文献で、審査請求が行われている特許に対して存在する。ただし、拒絶理由通知なしで特許権が成立したものについて当該データは存在しない。第二は、登録特許について特許公報上で公表されている引用文献情報である。登録特許のみであるが、拒絶理由通知なしで特許権が成立したものについても、審査官から見て重要な文献があると判断された場合には引用文献が存在する。しかし、整理標準化データにおいて1985年以降の特許公報におけるデータしか入力されていないという問題点がある。

すべての登録特許について、それぞれの引用情報が存在する特許の数を出願年別にグラフにしたものが図2である。拒絶査定に関する引用情報については、過去の文献（国内特許）を引用している特許の割合が年によって大きく上下している。これは制度改正や一部のデータの欠落、審査官の審査方針の変化によるものと考えられる。その一方で特許公報による引用情報として使えるのは1980年代後半以降のものに限られるという問題がある。従って、公開用のデータベースにおいては、この両者を提供するとともに、ここの引用情報がどちらの情報によるものなのかに関する識別記号を設け、分析ニーズにフレキシブルに対応するようにしている。

整理標準化データにおける引用情報は、主として特許審査の過程における拒絶理由となった過去の文献をデータベース化したものであり、米国特許における出願人による引用とは性格が異なる。しかしながら、累積的なイノベーションの実態を示す情報としては貴重である。NBER patent databaseにおける米国特許の引用情報との比較を行った結果によると日米である程度共通した引用パターンが見られた（注11）。

また、主として審査官による引用情報がベースとなっているということは、発明人はその情報を所与として発明を行ったのではないという可能性がある。その場合はイノベーションが累積的に行われているのではなく、たまたま同じ分野で重複して行われていたということになる。この点で米国特許における出願人による引用データは、累積的イノベーションやその過程における技術スピルオーバーをより直接的に示す指標であるということができる。ただし、米国特許の引用情報においても、出願者の特許弁護士やUSPTOにおける審査官が審査の過程で付与するものが少なくないことにも注意する必要がある。米国特許の引用情報は、2001年から出願人による引用と審査官による引用を区別して公表するようになった。その情報によると全体の四割強が審査官による引用となっている（Alacer and Gittelman, 2004）。また、米国特許の発明者に対するアンケート調査の結果によると、引用特許を発明の前に知っていたという割合は20％弱で、30％程度の発明者はそもそも知らないとしている（Jaffe, Trajtenberg and Forgarty, 2000）。従って、米国特許データにおいても、純粋な技術スピルオーバー効果以外の情報がかなり含まれていると見るべきである。

欧州特許庁の審査官のサーチレポートをベースとした引用情報は、整理標準化データの引用情報と類似性が高い。欧州特許庁の審査マニュアルによると当該特許の先行文献として適当な文献を必要最小限引用することとなっており、日本における審査ガイドラインと同様の考え方を示している。また、米欧の引用件数を比較すると欧州の件数は相当程度小さいことが分かっている（Michel and Bettels, 2001）。米国特許については、近年その出願件数が急増すると同時に、必要となる引用文献の取捨選択が厳密に行われておらずcitation inflationを起こしているという指摘もある（Hall, Jaffe and Trajtenberg, 2000）。このような点から、必要最小限の引用情報を盛り込む欧州における引用データの方が質の高い情報を提供しているとも考えることができる。

5 結語

物理学の巨人、ニールス・ボーアは、あるものは計測できるまでは存在しているとはいえない、と述べた。この言葉は計測の重要性をあらわすものとして広く引用されている。今後の知識社会において、知識の重要性はいっそう重要となるが、知識を計測することは容易ではない。そのなかで特許は技術知識についての豊かな知識の宝庫である。しかし、そのままでは計量的な分析に用いることは困難である。NBERのデータベースに代表されるように、世界的にこの豊かな技術知識の宝庫をより利用しやすいデータベースに編集し、これを用いて技術知識が生み出され利用されていくプロセスの分析に用いる試みがなされ、すでに大きな成果を挙げつつある。特許統計データは、われわれにあらたな分析の地平を開いてくれている。本論文で紹介したデータベースは日本の特許情報を包括的な計量分析のための初めてのデータベースである。このデータベースの作成にかかわったわれわれは、このデータベースが広く研究に利用され、日本の特許研究に貢献することを願っている。

注1 本データベースの開発にあたっては、多くの方々から多大の協力をいただいた。以下に記して、謝意を表したい。間中耕治、鈴木潤、内藤裕介、淺見節子、佐藤純子、長岡貞男、岡田羊祐、和田哲夫、玉田俊平太。

注2 Schmookler,J. *Invention and Economic Growth,* Harvard University Press,1966.

注3 「National Bureau of Economic Research：全米経済研究所」の略

注4 「European Patent Office：ヨーロッパ特許庁」の略

注5 「Organization for Economic Cooperation and Development：経済協力開発機構」の略

注6 2005年11月25日より公開の予定（URL: http://www.iip.or.jp/）。

注7 特許庁提供の整理標準化データをデータベース化する作業の詳細については内藤（2005）を参照されたい。

注8 整理標準化データ平成15年度第22回提供分までのデータを用いており、出願関係については2004年1月28日までの特許公報、特許庁における入力データについては2004年1月23日までの入力分が含まれている。

注9 出願人番号に関するデータクリーニング作業の詳細については、元橋(2005a)を参照。

注10 なおわが国でも2002年9月1日に先行技術文献情報開示制度が導入された。これにより、出願人は特許を受けようとする発明に関する先行技術のうち出願時に知っている文献名称名などの先行技術文献情報を明細書に記載しなければならない。ただし、罰則はない。

注11 これらの点については元橋（2005b）参照。

＜参考文献＞

- 元橋一之（2005a）「特許庁整理標準化データを用いた研究者用データベースの作成について」『特許データを用いた技術革新に関する研究』平成16年度特許庁研究事業　大学における知的財産権研究プロジェクト研究成果報告書2005年3月（以下、成果報告書とする。）
- 元橋一之（2005b）「整理標準化データにおける特許引用情報の活用可能性に関する調査研究」成果報告書
- 内藤祐介（2005）「研究用特許データベースの開発」　成果報告書
- Alcacer, J. and M. Gittelman, （2004）, How do I now what you know? Patent Examiners and the Generation of Patent Citations, mimeo
- Hall, B., Jaffe, A. and M. Trajtenberg （2000）, Market Value and Patent Citations: A First Look, NBER WP 7441
- Hall, B., Jaffe, A. and M. Trajtenberg （2001）, The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools, NBER WP 8498
- Jaffe, A. and M. Trajtenberg （2002）, *Patents, Citations and Innovations, A Window on the Knowledge Economy,* MIT Press
- Jaffe, A., Trajtenberg M. and M. Fogarty （2000）, The Meaning of Patent Citations: Report on the NBER/Case-Western Reserve Survey of Patentees, NBER WP 7631
- Michel, J. and B. Bettels （2001）, Patent Citation Analysis: A Closer Look at the Basic Input Data from Patent Search Report, *Scientometrics,* vol. 51, No. 1, pp. 185-201
- Schmookler,J *Invention and Economic Growth,* Harvard University Press,1966.