『IIPパテントデータベース』の開発と利用

中村, 健太

(Citation)
國民經濟雜誌,214(2):75-90

(Issue Date)
2016-08-10

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/E0040989

(URL)
https://hdl.handle.net/20.500.14094/E0040989

『IIP パテントデータベース』の開発と利用

中　　村　　健　　太

国民経済雑誌　第 214 巻　第 2 号　抜刷
平 成 28 年 8 月

『IIP パテントデータベース』の開発と利用

中　村　健　太

科学技術・イノベーション政策において，エビデンスに基づく議論，政策立案が重要である。また，エビデンスは信頼性の高いデータ，ツールによって導かれる必要がある。その意味では，現代のイノベーション研究において特許データは不可欠な存在である。2005年末には，我が国イノベーション・プロセスの評価・分析に資する基礎データとして研究用特許データベース『IIP 特許データベース』がリリースされた。また，2015年 7 月にメジャー・アップデート版である『IIP パテントデータベース2015年版』がリリースされた。同データベースには，1964年以降に出願公開・登録された特許（出願）約1,270万件が収録されている。本稿では，最新版での変更点，使用上の留意点，データベース開発を通じて理解が進んだ日本の特許データの特性などについて紹介する。

キーワード　　特許，IIP パテントデータベース，イノベーション

1　は　じ　め　に

科学技術・イノベーション政策の場において，エビデンスに基づく議論，政策立案の重要性が説かれて久しい。また，エビデンスは信頼性の高いデータ，ツールによって導かれる必要がある。その意味では，現代のイノベーション研究において特許データは不可欠な存在だと言える。

特許データは，一般に以下のような特徴を持つ。第一に，システマチックなデータ集積体系が挙げられる。我が国であれば，特許庁において一元的にデータが管理されている。第二に，特許制度には長い歴史があり，また世界的に広く採用されている。このことは，時系列および地域間の比較可能性を意味する。第三に，個々の特許に国際的に統一された技術分類である IPC（International Patent Classification: 国際特許分類）が付与されている。例えば，研究開発費のデータであれば，そのボリュームは分かっても技術分野は不明である。第四に，技術分野に偏りがなく全分野をカバーしている[1)]。第五に，引用関係から先行・後方技術の把握ができる。第六に，基本的に公開情報のため，データの入手が容易である。

これらの特徴は，元来特許情報に備わる属性であるが，イノベーション研究が進展し，エビデンスが見いだされるようになったのは，NBER で米国特許の研究用データベース

『NBER U.S. Patent Citations Data File』（以下，「NBER データ」）が公開されたことによる（Hall et al., 2001）。これまでも商用の特許データベースは存在したが，企業などにおける先行技術調査が主たる用途であり，大容量データによる統計的な分析は想定されていなかったのである。

当時，日本には研究用の特許データベースが存在しなかったため，日本企業に関する特許データは，NBER データから米国特許を抽出するという方法が採られていた。しかし，こうして得られた特許は重要発明や米国向けの発明に偏る可能性が高い。日本企業の行動，日本の特許制度，日本の技術動向を分析する際には，日本の特許データを用いるのが自然であり，我が国でも研究用特許データベースの整備が待たれていた。Sakakibara and Branstetter（2001, p. 86）に当時の日本における特許データ事情を表す興味深い記述があるので下に引用しておく。同論文は，日本の特許制度が改善多項制に移行したことの影響を分析したものであり，特許制度の強化が必ずしも研究開発インセンティブを高めないことを示した研究として有名である。なお，引用中の JAPIO とは，日本で最初の特許情報オンライン検索システムを開始した日本特許情報機構（Japan Patent Information Organization）のことである。

“This analysis also requires data on Japanese patenting at the firm level. Regrettably, such data are difficult to obtain and, relative to our U.S. data, extremely expensive. Despite the assistance of the staff of JAPIO, which provides the only practical electronic patent database in Japan, and despite their provision of a generous discount on the rates charged to commercial users, we were able to obtain only two series of patent data for each of our sample firms.”

NBER データは，米国で登録された特許のみを対象としている。そこで，日本においても同様のデータを整備しようとするプロジェクトが発足し，東京大学先端科学技術研究センターの後藤晃教授（所属および肩書きは当時のもの）を中心に多くの研究者によって進められた。同プロジェクトは，特許庁の『整理標準化データ』をソースとして，NBER データと同じく，研究や分析において重要な書誌情報に限って無料で提供しようとするものであり，『IIP 特許データベース』（以下，「IIP-DB」）として2005年末に当時の財団法人知的財産研究所（IIP）から最初のバージョンがリリースされた（後藤・元橋, 2005；鈴木・後藤, 2007；Goto and Motohashi, 2007）。これには，1964年 1 月から2004年 1 月までに出願公開・登録された特許（出願）約900万件が収録されていた。

2015年 7 月に IIP-DB のメジャー・アップデート版である『IIP パテントデータベース2015年版』（以下，「IIP-DB2015」）がリリースされた（中村・IIP パテントデータベース運営委員会，2015）。本稿では，最新版での変更点，使用上の留意点，データベース開発を通

じて理解が進んだ日本の特許データの特性などについて紹介する。

2　IIP-DB2015 の概要

『IIP パテントデータベース2015年版』は，2015年 7 月に公開された IIP 特許データベースのメジャー・アップデート版であり，本稿執筆時点（2016年 4 月）で最新のものである。データベースは，一般財団法人知的財産研究教育財団知的財産研究所のホームページから入手できる[2)]。

2. 1　データソース

データソースは，2013年度末までに特許庁から提供された整理標準化データであり，「1964000001」以降の出願番号を持ち，出願公開・登録された特許（出願）約1,270万件が収録されている[3)]。

特許庁では，出願された特許，実用新案，意匠，商標に関する書誌情報および経過情報を，SGML 形式あるいは XML 形式で整理標準化して提供している。これが『整理標準化データ』である（以下，特許に限定して説明を進める）。特許庁が提供する電子データとしては，他に公開特許公報や特許公報などの公報データがあるが，こちらが公開や登録時の固定されたデータであるのに対し，整理標準化データは，イベントの発生によって更新されていくデータである（鈴木・後藤，2007）。

整理標準化データで提供されるデータは，特許庁内の各マスタ・データベースにおいて発生した新規データおよび更新データであり，具体的には「出願マスタ」,「登録マスタ」,「サーチマスタ」,「引用文献マスタ」,「IPC8 版マスタ」,「審判マスタ」から構成される。提供対象は，公開特許公報，公表特許公報，再公表特許，公告特許公報，特許公報が発行された案件である。収録期間は1960年代から直近までであり，同データを用いることで，日本の特許について長期かつ詳細な情報を得ることができる。

整理標準化データでは，イベントの発生に応じて更新データ（差分ではなく全体）が提供されるため，通常，1 出願に対して複数（相当数）のレコードが発生する。IIP-DB2015 の作成時点ではのべ件数で約17,000万件のデータが存在した。SGML 形式および XML 形式のデータは，タグ情報を解析した後，リレーショナル・データベース化してある。ここから必要なデータを抽出することで IIP-DB2015 が作成される。なお，「原則として整理標準化データに忠実」にデータベースを作成した。したがって，整理標準化データ上に誤記がある場合，それを個別に修正することはしていない。

2.2 IIP-DB2015 の構成

IIP-DB2015 は，表 1 に示す 5 つのテーブル（ファイル）から構成される。各テーブルの内容は，次節で説明するが，基本的にはテーブル名から推測されるデータが含まれる。なお，IIP-DB2015 の「出願テーブル」は，出願日などの各種日付や技術分類，請求項数を含んでおり，かつての IIP-DB で「特許出願ファイル」,「特許登録ファイル」と呼ばれていたデータを足したものに近い。また，IIP-DB の「出願人ファイル」,「発明者ファイル」,「権利者ファイル」は，純粋に出願人名称などしか含んでおらず，それらのデータを「特許出願ファイル」や「特許登録ファイル」と接続するためのテーブルが別途用意されていた。今回のアップデートでは，接続用テーブルを使用するという以前の方法を廃し，表 1 の各テーブルに出願番号を持たせることによって,「出願人テーブル」から「引用テーブル」までを「出願テーブル」へ直接に接続できるよう改めた[4)]。ファイル形式は，すべてタブ切りテキストファイルであり，文字コードは UTF-8 によって記述されている（要解凍）。

表 1 テーブル一覧

	テーブル名	ファイル名	レコード数
1	出願テーブル	ap.txt	12,706,640
2	出願人テーブル	applicant.txt	13,700,370
3	発明者テーブル	inventor.txt	25,499,350
4	権利者テーブル	hr.txt	4,922,168
5	引用テーブル	cc.txt	20,036,172

2.3 提供データの範囲

2.1 で述べた通り，IIP-DB2015 は，2013年度末までに提供された整理標準化データから作成されている。また，整理標準化データにおける新規データの発生条件は，公報の発行であった。つまり，公報が発行されていない案件は，整理標準化データに現れないため，IIP-DB2015 にも収録されない。具体的には，①出願公開制度導入以前で公告に至らなかった出願，②出願公開前に何らかの理由で取り下げられた出願，③出願後18ヶ月を経過していないため未公開の出願などが収録されていない。

表 2 は，IIP-DB2015 を用いて出願年別の件数 (a) を示している。また，ベンチマークとして『特許行政年次報告書2015』（特許庁）に記載されている出願件数 (b) を併記した。両者は集計方法が異なるため（年次報告書には，取下なども含まれている），比率（a/b）は100％にならないものの，2011年までは90％前後で推移し，翌2012年には比率が57％にまで低下している。これは2012年後半に出願された特許が未だ公開に至っていないことによる。

逆に2013年の出願は，出願後18ヶ月未満にもかかわらず，36,000件ほど収録されている。

ただし，これらの特許は，出願から審査請求までの期間が短く，また，審査期間も短いものに偏っている可能性がある。特許が早期に成立する要因としては，当該発明の進歩性が極めて高い場合や，逆に，特許請求の範囲が狭い場合など様々な理由が考えられるが，それらの特許群が当該出願年における平均的なサンプルと見なせるかは明らかでない。このように，データソースの仕様上，最近の出願においてトランケーションやバイアスが発生する可能性があることを念頭に置いて分析期間を設定するなどの処置が必要である。

表 2　収録件数（出願テーブル）

出願年	IIP-DB2015 (a)	特許行政年次報告書 (b)	(a/b)
2008	351,065	391,002	90%
2009	316,680	348,596	91%
2010	310,416	344,598	90%
2011	296,561	342,610	87%
2012	196,193	342,796	57%
2013	36,666	328,436	11%
2014	2	325,989	0%

2.4　データの読み込み

IIP-DB2015 は，複数のタブ切りテキストファイルから構成される。テキスト形式なので，任意のソフトウェアでハンドリング可能であるが，データの読み込みに関してユーザーからしばしば質問を受けることがある。

一つは，出願人名称や住所など日本語の「文字化け」である。前述の通り，IIP-DB2015 は，文字コードに UTF-8 を採用している。そのため，UTF-8 を全くサポートしていない，あるいは，初期設定の文字コードが他のものになっているソフトウェアでは，文字化けが発生する可能性がある。例えば，バージョン13までの Stata では，文字化けが発生した（バージョン14はネイティブに UTF-8 をサポートしている）。文字コードを変換するツールはフリーでも各種存在するし，また，UTF-8 をサポートするテキスト・エディタで一度ファイルを読み込み，別の文字コードで保存し直すことでも文字化けを解消することができる。

もう一つは，ファイルの全体が読み込めないという現象である。表 1 に示したように，各テーブルは数百万以上のレコードを持つため，元のファイルを直接 Excel で処理するのは無理がある。Stata など統計パッケージ・ソフトの利用を推奨したい[5)]。また，IIP-DB2015 からの試みとして，各テーブルを出願番号で分割したファイルの提供も行っている。メモリ搭載量が少ない PC で処理を行う場合などは，こちらを利用することも一案である。このほか，

バージョン13以降の Stata ユーザーに見られる現象として，出願人テーブルや発明者テーブルを読み込むとメモリが不足するといったものがある。これは，一部の変数についてデータの型を変更することで解消できる。詳細は巻末注を参照されたい[6)]。

3　各テーブルの概要

本節では，IIP-DB2015 を構成する 5 つのテーブルについて内容を説明する。

3. 1　出願テーブル

出願テーブルは，各種番号，日付，IPC，請求項数を出願番号単位に記録したものである。具体的な収録内容は，表 3 の通りである。

表 3　出願テーブル

	ラベル	概　　要	備　　考
1	ida	出願番号	先頭 4 桁の出願年と下位 6 桁の連番からなる10桁の番号
2	adate	出願日（西暦）	YYYY-MM-DD 形式
3	sdate	審査請求日（西暦）	YYYY-MM-DD 形式
4	idr	登録番号	最大 7 桁の番号
5	rdate	登録日（西暦）	YYYY-MM-DD 形式
6	tdate	権利消滅日（西暦）	YYYY-MM-DD 形式
7	class 1	公開・公表の筆頭 IPC	セクション，クラス，サブクラスをまとめた 4 桁の記号
8	group 1	公開・公表の筆頭 IPC	メイングループ，セパレータ，サブグループをまとめた最大11桁の記号
9	class 2	公告・登録査定時の筆頭IPC	公開・公表の筆頭 IPC と同様
10	group 2	公告・登録査定時の筆頭IPC	公開・公表の筆頭 IPC と同様
11	claim 1	請求項数（出願時）	1999年以前に PCT から国内移行した案件の場合，真の請求項数にかかわらず整理標準化データ上の請求項数が 1 になっている場合がある
12	claim 2	請求項数（公告決定時）	
13	claim 3	請求項数（登録査定時）	

3. 1. 1　出願番号

出願番号は，IIP-DB2015 において最も重要かつ基本的な情報である。すべての出願には

一意に出願番号が付与されている。また，IIP-DB2015 では，各テーブルの出願番号が等しいことを条件としてテーブルの結合ができる。[7)]

特許データを扱う上で，出願番号の基本ルールを知っておくことは有益である。例えば，公開特許公報などでは，「特願平9-259」や「特願2000-357」といった出願番号が記載されている。これは，それぞれ平成 9 年の特許出願の259番，同様に2000年の357番を意味する。そこで，IIP-DB2015 では先頭 4 桁の出願年と下位 6 桁の連番からなる10桁の番号として出願番号を収録している。したがって前述の例は，IIP-DB2015 上ではそれぞれ「1997000259」，「2000000357」となる。

出願番号の下位 6 桁は，表 4 に示す番号体系を持つ。PCT 国際出願が日本に国内移行した場合，10桁表記の出願番号の 5 桁目（出願年の次の桁，下位 6 桁の先頭）が「5」か「6」であることは知っておくと便利である。[8)] 重要な発明であれば複数の国・地域での特許権取得を目指すことが多いため PCT ルートによる出願が選択される可能性がある。もっとも，外国への出願は PCT 以外の方法（いわゆるパリ・ルート出願）でも可能なため，国際的な特許出願の全体像を捕らえるためには別のデータが必要であるが，出願番号の外形的な特徴から発明の質に関する出願人の主観的評価の情報を読み取れる可能性がある点は重要である。

表 4　出願番号と出願の種類（1991年～）

	出願番号	出願の種類
1	000001～499999	国内・通常出願（電子出願）
2	500001～699999	PCT 出願（2000年～電子出願）
3	700001～799999	特許権存続期間延長出願
4	800001～999999	協定出願

出所：独立行政法人工業所有権情報・研修館（2008）「整理標準化データ仕様書 XML 編【第2.2版】コード表B0010」[9)]

3. 1. 2　日付データ

日付に関しては，出願日，審査請求日，登録日，権利消滅日を収録している。先願主義を採用する我が国では，発明のタイミングは出願日に近いと推測される。審査請求制度は，1971年に導入された。出願人あるいは第三者によって審査請求がなされた日が審査請求日であり，出願日とのラグ（審査請求ラグ）は，発明の価値に対する不確実性と正の相関があると考えられる。[10)] また，登録日と審査請求日（あるいは出願日）とのラグをグラント・ラグと言う。価値の高い発明ほど，出願人が早期成立を目指すため，ラグは短期化すると考えられる。ただし，グラント・ラグは，当該技術分野における審査の滞貨（バック・ログ）や，審査期間の短縮に係る政策などの影響を受ける可能性がある点は留意を要する。権利消滅日は，

ある時点での生存・保有している特許ストックを計算する場合に不可欠な情報である。また，権利維持期間（つまり出願日とのギャップ）は，技術の減耗率や発明の私的価値の推定に利用される。

3.1.3　技術分類

IPC は，世界の多くの特許公報で採用されている技術分類であり，第 7 版までは概ね 5 年ごとに改訂が行われていた。2006年 1 月からは第 8 版が採用されており，近年では（第 8 版の中で）年 1 回の改訂が行われている。IPC の構造は図 1 のようになっている。すなわち，全技術分野を 8 個の「セクション」に分け，各セクションを数個から数十個の「クラス」に分け，さらにそれは「サブクラス」→「メイングループ」→「サブグループ」へと細分される。

IIP-DB2015 に含まれる IPC コードは，「公開・公表の筆頭 IPC」と「公告・登録査定時の筆頭 IPC」である。特許出願に係る発明が複数の技術によって構成される場合，それぞれの技術に対応する IPC（メイングループあるいはサブグループ）が複数付与されるが，その中で最も中心的な技術が筆頭に表示される。これが筆頭 IPC である。

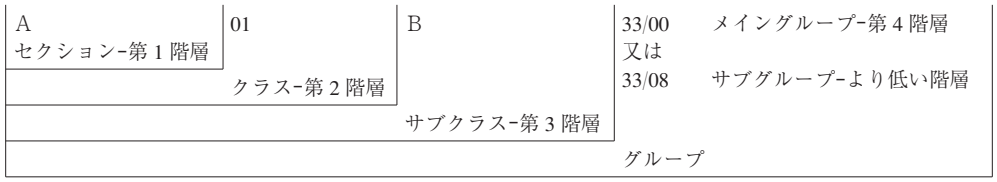
IPC のデータに関する留意点を 3 つ挙げておこう。第一に，「公開・公表の IPC」と「公告・登録査定時の IPC」の違いであるが，前者は特許審査の準備として出願公開の前に審査官や外部の登録調査機関によって付与された IPC，後者は公告ないし登録査定された特許請求の範囲について審査官（あるいは審判官）が付与した分類である。したがって審査段階で補正がなされた場合などは，２つの IPC コードが一致しないことがある。なお，出願公開制度（1971年から）以前は，「公告・登録査定時の IPC」のみ記入されている。

第二に，IPC の改訂についてである。IPC は技術進歩や出願動向を考慮して定期的に改訂されているため，同一の IPC コードであっても，改訂によってカバーする技術範囲が変化することがある。したがって，比較的細かな技術分類を用いて出願件数を時系列で集計する場合などでは，IPC 改訂による悪影響（ノイズ）が生じていないことを確認することは重要である[11]。なお，２つの筆頭 IPC は，整理標準化データの仕様上，IPC の第 2 版から第 7 版にしたがっている[12]。

第三に，「より粗い」技術分類についてである。IPC は非常に詳細な技術分野の情報を提供するが，例えば，マクロの技術動向を眺めるような用途ではサブクラス・レベルでも細かすぎる可能性がある。こうした場合の対処法には以下のものがある。

①サブクラスの文字列（アルファベット 1 文字＋数字 2 文字＋アルファベット 1 文字）からセクション（アルファベット 1 文字）あるいはクラス・レベル（アルファベット 1 文字＋数字 2 文字）を抽出する。このとき，決してアルファベット 1 文字＋数字 1 文字だけを抽出

図 1　IPC の概要



出所：特許庁（2015）「国際特許分類（2015年バージョン）（仮訳）指針」, p. 5.

してはいけない（図 1 参照）。

②WIPO（World Intellectual Property Organization: 世界知的所有権機構）が提供しているコンコーダンス・テーブル（The WIPO technology concordance table）を用いる[13]。同コンコーダンスは，IPC を WIPO の公式統計で用いられている技術分類に変換するためのものである（表 5 ）。IIP-DB2015 の IPC をコンコーダンスに通すことによって，WIPO の35分類が得られる。

表 5　IPC8 -Technology Concordance

Field number	Sector	Field
1- 8	Electrical engineering	Electrical machinery, apparatus, energy; Audio-visual technology; Telecommunications; Digital communication; Basic communication processes; Computer technology; IT methods for management; Semiconductors
9-13	Instruments	Optics; Measurement; Analysis of biological materials; Control; Medical technology
14-24	Chemistry	Organic fine chemistry; Biotechnology; Pharmaceuticals; Macromolecular chemistry, polymers; Food chemistry; Basic materials chemistry; Materials, metallurgy; Surface technology, coating; Micro-structural and nano-technology; Chemical engineering; Environmental technology
25-32	Mechanical engineering	Handling; Machine tools; Engines, pumps, turbines; Textile and paper machines; Other special machines; Thermal processes and apparatus; Mechanical elements; Transport
33-35	Other fields	Furniture, games; Other consumer goods; Civil engineering

3. 1. 4　請求項数

請求項の数は，出願人が保護を期待する発明の数を表しており，特許の価値指標の 1 つと考えられている。ここでの留意点は，2 つである。第一に，我が国では，1988年の改善多項

制の導入によって 1 出願に単一性の要件を満たす複数の発明を包含できるようになった。つまり，制度導入以前の特許について請求項を議論してもあまり意味がない。第二に，整理標準化データでは，PCT 出願で出願番号が「2000000001」より小さい出願において，真の請求項数にかかわらず請求項数が 1 になっていることがある。したがって，請求項数は2000年以降の出願についてのみ利用可能であると考えて良いだろう[14)]。

3. 2 出願人テーブル

出願人テーブルは，出願人情報を出願番号・記載順序単位（ida_seq）に記録したものであり，初期時点（出願により近い時点）の情報を収録している。出願後に出願人の名称が変更される場合や，特許を受ける権利が譲渡された場合，出願人情報に更新が発生する。ただし，研究開発主体の情報として出願人情報を利用することが多いことから，権利譲渡などの影響を受けない初期時点の出願人情報を収録している。したがって，原則として当該テーブルの情報は，将来の IIP-DB のアップデートに対して安定的である。具体的な収録内容は，表 6 の通りである。

出願人名称および住所は，原則として整理標準化データの情報をそのまま収録している。そのため，同一出願人であっても表記揺れが存在する場合がある。

表 6 出願人テーブル

	ラベル	概 要	備 考
1	ida	出願番号	
2	seq	出願人記載順序	
3	ida_seq	出願番号＋記載順序	出願番号（ida）と記載順序（seq）をアンダーバー（”_”）で接続したもの（記載順序は 3 桁表記）
4	name	出願人名称	
5	address	出願人住所	
6	idname	出願人コード	特許庁から付与された ID コード
7	country_pref	国県コード	国コードあるいは県コード（国コードは付録 1，県コードは付録 2 参照）
8	kohokan	個法官コード	個人(1)，法人(2)，官庁(3)，その他(9)の区別

3. 2. 1 出願人コード，国県コード，個法官コード

出願人コードは，特許庁から付与される ID コードであるが，いくつか注意点がある。第一に，出願人コードは1990年12月以降すべての出願人に付与されている。それ以前は，大規

模な出願を行っている一部企業のみコードが与えられていた。第二に，コードは，時系列的に改訂されることがある。第三に，同一出願人に複数のコードが付与されることがある（異なる事業所からの出願などの場合）。第四に，国県コードおよび個法官コードは，基本的には出願人コードが割り当てられなかった出願人に対して記入されている。したがって，出願人コードのみを頼りに名寄せを行った場合，取りこぼしが生じることになる（名称や住所の利用も必要である）。また，国県コードや個法官コードが適用可能な期間は1990年12月以前に限定される。ただし，1990年以前の住所情報と国県コードを参考にして，それ以降の出願に対してコードを付与することは可能である。

3.2.2 NISTEP 企業名辞書

これまで述べてきたように，出願人名称や住所には表記揺れが存在し，また，出願人コードも企業ごとの統一が図られていない。これらの点から想像が難くないように，従来，出願人の名寄せは，特許データを利用する上で深刻かつ手間のかかる障害であった。しかし，現在では文部科学省科学技術・学術政策研究所（NISTEP）の『NISTEP 企業名辞書』および IIP-DB2015 と企業名辞書の接続テーブルが公開されたことで名寄せの問題は相当部分解消した[15]。企業名辞書は，①特許出願数累積100件以上，②株式上場企業，③特許出願数の伸び率大の 3 条件のいずれかに合致する企業約6,500社（とその改組に関する情報）を含んでいる。また，接続テーブルには，企業名辞書における企業番号と出願テーブルの ida_seq の組が収録されている（つまり，企業名辞書収録企業に関する名寄せ情報が利用できる）。

3.3 発明者テーブル

発明者テーブルは，発明者情報を出願番号・記載順序単位（ida_seq）に記録したものであり，初期時点（出願により近い時点）の情報を収録している。したがって，原則として当該テーブルの情報は，将来の IIP-DB のアップデートに対して安定的である。具体的な収録

表 7　発明者テーブル

	ラベル	概　要	備　　考
1	ida	出願番号	
2	seq	発明者記載順序	
3	ida_seq	出願番号＋記載順序	出願番号（ida）と記載順序（seq）をアンダーバー（”_”）で接続したもの（記載順序は 3 桁表記）
4	name	発明者名称	
5	address	発明者住所	

内容は，表 7 の通りである。

発明者情報の用途は様々であるが，比較的シンプルなものとしては，各特許の発明者数をカウントし，当該発明の開発に係るインプットの代理変数とする方法がある。また，発明者の住所には企業名が含まれることがあり，その企業名と出願人名を比較すれば，当該発明者が社内の研究者か，社外の研究者かを識別できるかもしれない[16)]。このように，発明者の氏名や住所をテキスト処理することで，社外との共同研究開発の頻度や効果，共同研究の成否と距離の関係，研究者のモビリティーと生産性の関係などの分析への応用があり得る。

3.4　権利者テーブル

権利者テーブルは，登録済みの特許について，権利者の情報を出願番号・記載順序単位（ida_seq）に記録したものであり，整理標準化データから得られる最新の情報（権利者情報の変更が発生した場合は，更新された権利者情報）を収録している。収録内容は，表 8 の通りである。

表 8　権利者テーブル

	ラベル	概　要	備　　考
1	ida	出願番号	
2	seq	権利者記載順序	
3	ida_seq	出願番号＋記載順序	出願番号（ida）と記載順序（seq）をアンダーバー（”_”）で接続したもの（記載順序は 3 桁表記）
4	name	権利者名称	
5	address	権利者住所	
6	idname	権利者コード	特許庁から付与された ID コード

3.5　引用テーブル

引用テーブルは，審査官引用を引用特許の出願番号・被引用特許の出願番号の対応関係で記録したものである（表 9 ）。つまり IIP-DB2015 で提供される引用データは，発明者引用（特許明細書の中で発明者が引用する先行文献）ではない[17)]。また，日本特許と日本特許の引用関係に限定しているため，被特許文献の引用や外国特許の引用は含まれない。ただし，審査官引用は日本の特許が大部分であり，引用の対応関係を前記のものに限定することは必ずしも大きな問題にはならないと考えられる。

IIP-DB2015 の引用情報は審査官引用である。発明者引用の場合，重要な発明がより多く引用されるという関係は，直感的に理解しやすい。しかしながら，審査官引用は基本的には

拒絶理由を提示するための引用であり，被引用文献の選択に技術的有用性が関与してくるとは考えにくい。そうだとすれば，審査官による引用回数が多い特許が存在した場合，その現象はどのように解釈すれば良いのだろうか。この点について，山田（2015）は，審査官引用が多いことは，偶然の重複発明を表しているのではなく，先行技術が有望であることを理解した上で，意図的なフォローアップが行われた結果であると解釈するのが妥当であり，ゆえに審査官引用の頻度は，特許の価値指標になり得るとしている。

最後に引用データの連続性について重要な事柄を指摘する。IIP-DB2015 の引用データがいわゆる審査官引用であることはすでに述べた。元来，整理標準化データからは，引用の目的（引用種別）や引用が行われた日（起案日）の情報が入手できるが，実は，2006年の 4 月を境に収録される引用種別の範囲が拡大された。その結果として，整理標準化データに含まれる引用の件数が大幅に増加した。特に，拒絶理由通知書における「先行技術文献調査の記録」欄の引用文献が追加されたことの影響は大きい。これらの文献は，拒絶理由を構成するものではないが，出願人にとって補正の際に参考になるなど，有用と思われる先行技術がある場合に引用される。

以上の仕様変更（元々は特許庁の内部のシステム変更に起因する）は，2006年をまたいで引用データを用いた分析を行う際に重大な影響をもたらす可能性がある。例えば，2006年以前と以後で後方引用件数を比較し，それが増えている場合，先行技術からの知識フローが増加したと言えるだろうか（単に，収録される引用の範囲が拡大した影響かもしれない）。通常，こうした分析では，異時点間で引用データに質的な変化があると具合が悪い。そこで，IIP-DB2015 では拒絶理由通知の根拠として審査官が引用した特許を特定できるようにした（データのラベルは reason）。拒絶理由通知での引用は，引用種別の中でも主要なものである。また，前記の種別に関する引用は全期間で得られるため，（分析の目的やデータの加工方法にもよるが）拒絶理由通知での引用に限定してデータを利用すれば，仕様変更の影響を受けにくい。

表 9　引用テーブル

	ラベル	概　要	備　　考
1	citing	引用特許の出願番号	
2	cited	被引用特許の出願番号	
3	reason	引用タイプ	被引用特許が拒絶理由通知の根拠として引用された場合に 1，その他は空欄
4	search	引用タイプ	被引用特許が「先行技術文献調査結果の記録」においてのみ引用された場合に 1，その他は空欄

4　お　わ　り　に

本稿では，『IIP パテントデータベース2015年版』について，旧データベースからの変更点，使用上の留意点，データベース開発を通じて理解が進んだ日本の特許データの特性などについて解説した。データベースが広く利用され，イノベーション研究，またそれに基づく政策の議論が深化することを切に望んでいる。

今回はデータ自身の特性に重点を置いて説明を行った。また，紙面の都合上，意図的に説明を省いた部分もある。ここでは 2 点例示しておく。一つは，特許データを変数化して利用する目的，加工の方法，その理論的背景についてである。NBER データの登場以降，特許指標に関する研究は相当数の蓄積がある。それらを定期的に整理することは有益であろう。もう一つは，他の特許データベース，特に欧州特許庁（EPO）の『Worldwide Patent Statistical Database（PATSTAT）』との関係である。PATSTAT を利用できる環境においては，IIP-DB2015 との接続や使い分けも重要な課題になるだろう。これらの点については，リクエストがあれば解説の機会を設けたい。

注

＊　本稿の内容は執筆者個人の見解に基づくものであり，知的財産研究所および IIP パテントデータベース運営委員会の公式見解ではない。

1）もっとも，産業や技術分野によって特許の出願性向が異なる点は留意が必要である。

2）一般財団法人知的財産研究教育財団知的財産研究所のホームページ（https://www.iip.or.jp/patentdb/index.html 2016年 4 月10日アクセス）。なお，財団法人知的財産研究所は，一般財団法人に改組後，2016年 4 月に一般社団法人知的財産教育協会と合併し，一般財団法人知的財産研究教育財団となった。

3）非常に古い特許の一部で出願日の情報が得られないことがあるため，データベース収録の条件を出願日から出願番号に変更しているが，旧データベースで採用されていた基準とほとんど差はない。整理標準化データには，1964年より前のデータも一部含まれるが，必ずしも体系的にデータが得られないという理由から，これらについては IIP-DB2015 への収録を見送っている。

4）旧データベースが接続テーブルを含む構造になっていたのは，当時のマシンパワーやインターネット環境を考慮してファイルのサイズを小さくすることを目指したことによる。他方で，出願人ファイルなどのサイズを小さくするためにデータを正規化すると，異なるバージョンのデータベース間での整合性がとりにくくなるという問題があった。

5）もちろん，MySQL などのデータベース管理システムを用いてもよい。

6）バージョン13以降の Stata では，最大長が2045バイト以下の文字列である変数に対して，最大長分のメモリを事前に割り当てる。そのため，出願人や発明者の名称や住所に長い文字列が含まれていると，多くのメモリが消費され，ファイル全体が読み込めないことがある。そこで，テーブル分割版のファイルを使用し，ファイルを読み込んだ上で，非常に長い文字列を含む変数の型

を「strL」に変換し，必要に応じて分割されていたファイルを結合すればよい。

7）ただし引用テーブルは，引用特許および被引用特許の出願番号を含む。

8）特許協力条約（PCT: Patent Cooperation Treaty）に基づく国際出願とは，「ひとつの出願願書を条約に従って提出することによって，PCT 加盟国であるすべての国に同時に出願したことと同じ効果を与える出願制度」（https://www.jpo.go.jp/seido/s_tokkyo/kokusai1.htm 2016年 3 月25日アクセス）である。

9）番号体系は，年代によって若干の変更がある。「整理標準化データ仕様書 XML 編【第2.2版】コード表 B0010」，（http://www.inpit.go.jp/content/100030130.pdf 2016年 3 月25日アクセス）.

10）複数回審査請求がなされた場合は，請求の受理・却下にかかわらず，最初の審査請求日を採用している。

11）版を統一して IPC の再付与をすれば，上で述べた問題は回避できるはずである。EPO（欧州特許庁）の『Master Classification Database』では，過去の版の IPC は第 8 版（の最新版）に変換して収録されている。

12）つまり，概ね2000年以降の出願は，IPC 第 7 版にしたがっている。

13）コンコーダンスは，WIPO のホームページ（http://www.wipo.int/ipstats/en/ 2016年 3 月25日アクセス）で入手できる。Schmoch（2008）参照。

14）もし公報データが利用可能であれば，「特許請求の範囲」のテキストデータから請求項数を作成することができる。

15）詳細は，「NISTEP 企業名辞書（Ver.2015.1）利用マニュアル」を参照されたい。データおよびマニュアルは，NISTEP のホームページ（http://www.nistep.go.jp/research/scisip/rd-and-innovation-on-industry 2016年 3 月25日アクセス）で入手できる。

16）ただし，発明者住所の住所が自宅になっている場合もあるため，この方法が利用できないケースもある。

17）発明者引用のデータは，株式会社人工生命研究所の『公報データベース』で提供されている。

参 考 文 献

Goto, A. and Motohashi, K.（2007）“Construction of a Japanese Patent Database and a First Look at Japanese Patenting Activities,” *Research Policy*, Vol. 36, No. 9, pp. 1431-1442.

Hall, B. H., Jaffe, A. B., and Tratjenberg, M.（2001）“The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools,” NBER Working Paper, 8498.

Sakakibara, M. and Branstetter, L.（2001）“Do Stronger Patents Reduce More Innovation? Evidence from the 1998 Japanese Patent Law Reforms,” *RAND Journal of Economics*, Vol. 32, No. 1, pp. 77-100.

Schmoch, U.（2008）“Concept of a Technology Classification for Country Comparisons,” Final Report to the World Intellectual Property Organisation（WIPO）.

後藤晃・元橋一之（2005）「特許データベースの開発とイノベーション研究」，『知財研フォーラム』，Vol. 63, pp. 43-49.

鈴木潤・後藤晃（2007）「日本の特許データを用いたイノベーション研究について」，『日本知財学会誌』，Vol. 3, No. 3, pp. 17-30.

中村健太・IIP パテントデータベース運営委員会（2015）『IIP パテントデータベース・ユーザーマ

ニュアル』，一般財団法人知的財産研究所.
山田節夫（2015）『特許政策の経済学：理論と実証』，同分舘出版.