話題 WADAI

ライフサイエンス系データベースシステムの現状と将来展望

小河邦雄
Kunio OGAWA
大正製薬(株)総研研究システム部

1 はじめに

近年の科学情報量の急激な増加や学際的な研究展開，さらに経済的な事情によって，必要な情報にアクセスできない状況が進みつつある．それは，個人の情報スキルだけでなく，所属する機関の情報環境による情報デバイド(格差)も原因である．本稿では，薬学研究にとって最も大切なライフサイエンス関連の論文を探すためのデータベースについて，その歴史的な側面を踏まえた現状の問題点，将来展望を考えてみたい．

2 ライフサイエンスデータベース

薬学研究には学際的な多くの分野の研究情報が必要であり，パラダイム論を提唱した科学史家クーン(Kuhn T. S.)の言う近代科学の特徴「集団的営為」[1]で進められるとすれば，良好な情報流通は最も重要な課題であろう．さらに，伝統薬など数千年の経験知の積み重ねが基になる部分もあり，研究された成果についてはすべての研究者が自由にアクセスできる環境がその発展のためには欠かせない．しかし，実際には情報格差が生じており，集団的営為であるはずの研究を，必要な情報パズルが欠けた中で進めなければいけないとしたら，大きな問題である．特に，薬学研究には情報検索が不可欠であるが，研究者は所属機関が契約するデータベースしか使えない．例外として，米国の税金で作られるPubMed等もあるが，他のメジャーなEMBASE，BIOSIS，DDFや，化学や特許情報も含んだCA(Chemical Abstracts)，科学全般のScisearchの6個は有料である．これらは40年以上にわたって論文に抄録，独自の索引をつけて整理，蓄積してきた近代科学の知恵袋と言え，これらなくしては薬学研究も進まなかった．ライフサイエンス関連のデータベースとしては，他に遺伝子情報など実際のデータを収録，解析するものも数多く存在するが，それらは最近の紹介記事に譲り，[2] 本稿では文献データベースを中心に説明する．

3 主なライフサイエンスデータベースの起源

1836年にワシントンで設立された米国国立医学図書館は，軍事総督部の小規模な蔵書から出発した．そして，1879年に医学研究のために創刊された論文二次資料誌「Index Medicus」がMEDLINEのルーツである．軍の機関が始めたのは，南北戦争時に疾病で多数の死亡者が出たことや資金などが関係している．一方，MEDLINEとよく比較されるEMBASEは1947年にオランダで設立された非営利団体Excerpta Medica Foundationが創刊した医学抄録誌が起源であり，1971年にエルゼビアが親会社となり，雑誌出版とそのデータベースを持つ巨大情報企業へと至る．BIOSISは1926年に米国フィラデルフィアで設立された非営利団体Biological Abstracts, Inc.が起源であり，生物学と学際的分野(医学，薬学等)の両方を含む情報を提供してきたが，2004年に同じフィラデルフィアのトムソンISIの傘下に入った．DDF

(Derwent Drug File)は欧州製薬企業の Roche と Sandoz が 1920 年から始めた論文抄録と索引の共同作成が起源であり，1964 年に英国の Derwent 社に委嘱し，その後 Derwent 社はトムソンの傘下に入った．このように，ライフサイエンスデータベースの起源は非営利的団体であったが，近年の医療ニーズの高まりと共に規模が拡大する課程で戦略的な情報資産の側面も強まった．

4 文献検索システムの開発

冊子体資料を検索するにはコンピュータの登場が必要で，1960 年代に最初の文献検索が実現した．これは「Index Medicus」等の論文二次資料の編集・印刷にコンピュータが導入された結果，副産物としてデータが蓄積され，データベースとなった．1957 年にソ連が人類初の人工衛星スプートニクの打ち上げに成功し，これにショックを受けた米国は，翌年 NASA を設立し，研究開発に不可欠であった情報検索システムについてもロッキード社と共同で開発し，1967 年に完成した．その後，NASA の情報部門が Dialog として独立し，1972 年にオンラインサービスを開始した．そして遅れること 10 年近く経った 1980 年代に，日本の悲願であった海外のオンラインデータベースの使用が可能になり，研究者は冊子体を調査する苦労から解放された．しかし，当初は高額な検索費用の問題もあり検索専門家が対応した．主なシステムは米国の DIALOG，STN，ORBIT，BRS，フランスの QUESTEL/DARC，スイスの DataStar，そして日本の JOIS であった．

5 エンドユーザー検索の普及と問題点

1980 年代に質，量ともに発展したオンラインデータベースは，1990 年代に研究者自身で検索する「エンドユーザー検索」が始まった．1990 年代後半にはインターネットが普及し始め，データベース作成機関が直接，利用者へ情報サービスを行うことが可能になった．これらはメニュー形式の使いやすいシステムであったが，個々の異なるシステムを覚える必要があったり，複数のデータベースで大量の重複した情報が得られるなどの欠点もあった．また利用形態等の契約も異なり，管理負担も大きくなった．2000 年代になると，日本の JST(科学技術振興機構)が作成する JDream が比較的安価な固定費のサービスも始まり，国内科学誌の調査環境は大きく向上した．しかし，海外のデータベースについては情報企業の吸収合併が進み，トムソンやエルゼビアなどの巨大情報企業が生まれ，電子ジャーナルを含めた情報製品の数の増大と価格の上昇が進み，情報投資をどのように行うかが非常に難しい問題となってきた．このように検索システムは大きな発展を見せたが，コンテンツは情報企業にとっての貴重な戦略的情報資源となり，利用者はそれらの動きに翻弄されることになった．

6 主な文献データベースと特徴の概略

主なライフサイエンス系文献データベース 5 個の特徴を，表 1 にまとめた．米国の MEDLINE とオランダの EMBASE が医学・薬学分野で重要であるが，収録雑誌の重複は 6 割程度しかない．さらに索引方針が異なり，EMBASE は薬剤名を多数付与するなど，実際に検索すると収録雑誌の差以上に異なる．[3] その他の BIOSIS，DDF，CA も収録雑誌，索引システムに特徴があるために，ライフサイエンス分野の網羅的な調査を行う場合は，STN などの検索システムを使用して，5 個のデータベースを同時に検索し，その結果を重複除去する必要がある．ただ，検索主題が複雑な場合は，それぞれのデータベースに固有のキーワードと検索式を使用する必要があり簡単ではない．製薬企業における新規性調査や安全性調査などは網羅

表1　主なライフサイエンス系文献データベースの特徴

	MEDLINE	EMBASE	BIOSIS	DDF	CA
収録情報	生物医学及び薬学，歯科学，看護学などの幅広い文献情報	生物医学及び薬学領域の文献情報	生物及び生物医学分野の広範囲な文献情報	医薬品の合成，評価，製造，使用などの文献，会議録情報	化学，生化学，化学工学分野を中心とした文献，特許情報
収録期間	1947年～	1947年～	1926年～	1964年～	1840年～
収録件数	1,840万件	1,310万件	1,980万件	130万件	3,170万件
特許	×	×	△	×	○
統制語	○	○	○	○	○
作成機関	米国医学図書館	エルゼビア	トムソン・ロイター	トムソン・ロイター	CAS(米国化学会)
特徴	MeSHというシソーラスを持ち，最新のキーワードで過去に遡った検索ができる．	医薬品のキーワード付与が充実し，欧州の文献に強い．	会議資料を多数収録．概念コードによる研究分野の特定や生物系統分類コードを持つ．	明確な方針に従って論文を厳選し，薬剤ごとにキーワードがリンクされ，第三者抄録を持つ．	CAS登録番号による化学構造やタンパク配列の検索結果から文献が検索でき，物質ごとに詳細な索引を付与．

性が重要となるため，情報部門が責任を持って行っている．

7　増え続ける情報量

2009年の論文数は上記の5データベースの合計でのべ442万件もあり，1日換算で1.2万件となる．MEDLINE単独でも1日2,000件程で，1人でこれらをすべて見ることは不可能であるため，キーワードを登録し，自分のテーマに関係した論文を選択するシステムが必要となる．無料のPubMedでもキーワード登録ができるが，STN等で登録すれば，複数のデータベースを検索して重複除去した結果をメールで受け取ることも可能である．もちろん，仕事の内容や目的による違いで，詳細情報が必要な場合と主要な情報だけが欲しい場合とで情報源や情報の形態が異なるので，それに対応したシステムを選択する必要がある．そのため，それぞれの組織に最適な情報源を導入し，その適切な利用を支援する図書や情報の担当者の役割が重要となる．

8　ライフサイエンスデータベースの新しい流れと当社の対応

1980年代以降のライフサイエンス系文献データベースの中で系統的な階層構造を持った索引を付与するのはMEDLINE，EMBASE，BIOSISの3つである(図1)．1997年にMEDLINEがPubMedとして公開され，世界のメジャーな医学関連雑誌を無料で検索できるようになったことで，創薬研究においても標的分子に関する最新の基礎医学情報調査等に使われた．PubMedは，遺伝子情報や電子ジャーナルへのリンクなどにも優れ，従来の有料のMEDLINEの使用は激減することとなった．

2000年以降はデータベースのWeb版化が加速し，当社もエルゼビアのScopusを2006年に導入した．これはトムソン・ロイター(以下，トムソン)のWeb of Scienceを意識した製品であり，被引用の多い注目論文を簡単にリスト化できる．[4] エルゼビアは，EMBASEについても同名のWeb版に注力しており，オンライン版より詳細な検索が可能で，MEDLINEのデータも重複除去されて入っているなど，薬学研究についてはこれ単独でかなり網羅的な検索が可能となっている．これらはサイエンス・ダイレクトの電子ジャーナルとも連携して効率的な調査が可能となり，同社の電子ジャーナルの価値をさらに高めている．一方，トムソンのサイエンス部門はWPIなどの特許データベースで評価が高く，文献についてはDDFを作成していたが，年間収録数は5万件程度と製薬研究に特化していた．しかし，1992年に現在のWeb of

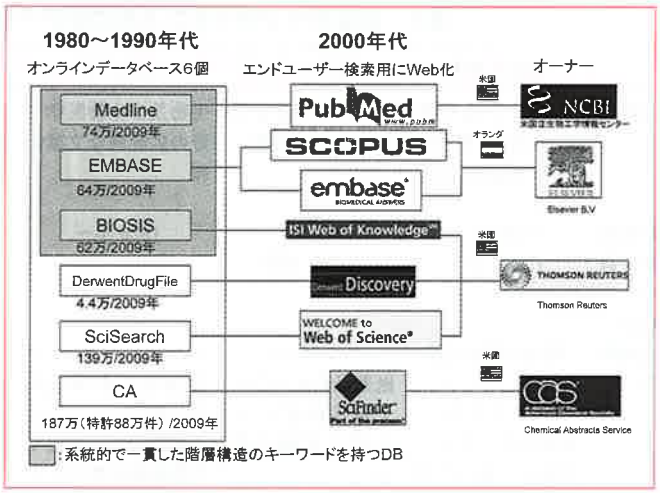

図1　Web化されたライフサイエンスデータベース

Science を作成していた米国の ISI を買収し，2004 年には BIOSIS を傘下にしたことにより，Web of Knowledge のプラットフォームで広範囲の文献情報提供を始めた．また，CA は統制された索引や自然言語に近い索引などを有し，柔軟な検索が可能である．化合物やタンパク質などの物質検索では他のデータベースを凌駕する網羅性を持っており，さらに生化学や前臨床試験などライフサイエンス関連の文献も多く含まれている．

1995 年に発表された SciFinder はインターネット経由で CA が検索できるエンドユーザー向けの製品であり，化合物の構造検索も可能であるため，研究者自身で新規性調査が可能となった．そのため，既知と判明した場合は調査部門への依頼が不要となる場合もあり，調査が最適化された．当社では 1999 年に化学部門に導入し，2010 年からはエンタープライズ契約を行い，研究員全員が固定費で使用できるようになった．検索結果に対して各種の解析が簡単にできるので，たとえ検索結果が多い場合でも目的の文献の集合を絞り込みやすく，MEDLINE との同時検索と重複除去も可能など，利用者の検索行動を良く考えたシステムといえる．このように，MEDLINE に関しては他のデータベースと併用して使用できる製品は多いが，複数のデータベースを一度に固定費で検索するエンドユーザー検索向けのツールは OvidSP[5] などのアグリゲーションサービス*にも適している．

9　創薬研究に必要な情報投資

製薬企業が創薬研究に使用する外部データベースへの投資は，研究部門だけでも 2～3 億円は必要と思われるが，トムソンやエルゼビア等の情報企業が勧める新製品をすべて導入していると，この数倍近くの費用が必要となる．この中で文献データベースは数千万円であり，その他は特許，化合物関連のデータベースであるが，最も額が大きいのは医薬研究開発関連のデータベースである．これは創薬に必要な情報を文献，特許，学会，Web 等から評価・選択した製薬企業の製品開発動向やパイプラインを調査するための統合システムである．トムソン，Prous，Adis，IMS，Informa healthcare の 5 社に歴史がある．中でもトムソンは資本力を背景に，競合する会社を買収していった(図 2)．

従来から特許を評価し，新規な Drug 情報を蓄積していた 3 企業の中で，2002 年に Current

＊　アグリゲーションサービスについての用語解説は，1141 頁参照．

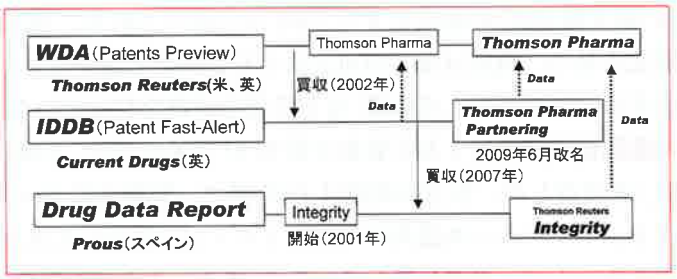

図2　特許由来の創薬情報データベースの変遷

Drugsを買収してIDDBを入手し，さらに2007年にProusを買収してIntegrityを手に入れた．これによって，医薬研究開発統合データベースの市場はトムソンの独占に近い状況となった．トムソンは手に入れたその他の多くのコンテンツと一緒に，1つのプラットフォームでサービスする理想的なシステムを目指している．しかし，コンテンツの種類とボリュームの大きさゆえにスピードの改善や価格の安定化など利用者に配慮したサービスが求められる．また，データベースの品質は個々の情報源を精査し，内容を的確に理解し，さらにキーワードを付与し情報を抽出する索引者(アナリスト)のレベルに依存する．特に統合データベースの場合はここが重要であり，この部分の質の維持について期待するとともに，利用者としても注目する必要がある．

10　ライフサイエンス系のデータベースの活用パターン

主要なMEDLINE，EMBASE，CAを中心に活用パターンを考えた．研究所，大学では，SciFinderでMEDLINEとCAを一度に検索し，重複除去するのが基本で，生物系はEMBASE Web版を使用して，重複除去されたEMBASEとMEDLINEを一度に検索するのが効率が良い．またOVIDを使用して，MEDLINE，EMBASE，BIOSISを一度に使用して重複除去する方法もある．DIALOGやDATASTRのオンラインシステムを使用して，MEDLINE，EMBASE，CAを重複除去する場合は従量制となり，STNシステムを使用した同様の方法は情報部門向きであるが，使用料は高額となる．新規分野の検討などはPubMedで検索し，SCOPUSかWOSで引用検索する方法が考えられる．無料のPubMed，Google scholar，SCIRUSで検索するのは，とりあえずというところか．このように網羅的な調査をするためには，複数のデータベースを調査する必要があるが，データベースを情報解析して新たな知識を得るデータマイニングには無料のPubMedの情報がデータセットとして使われることが多い．しかし，ライフサイエンス分野の情報量としては十分でない点があったが，最近，EMBASEの情報もデータ解析用に販売するようになった．EMBASE，CAはMEDLINEに収録されない多くの情報と独自の索引を持つため，情報解析データセットとしての有用性は高いと考える．化合物や標的分子などの新たな機能を情報から発見することが期待される．

11　薬学研究情報の研究会

データベースへの投資は高額であり，それを導入し情報支援を行う部門の責任は大きい．これらは図書館や情報支援の部門で行われているが，ここに所属する担当者はインフォプロ(information professional)[6]として，薬学情報に共通する問題点の情報交換や研修の場が必要となる．日本薬学図書館協議会や日本製薬情報協議会は数十年の歴史を持つ研究会であり，会員の情報スキルの向上を支援している．

日本薬学図書館協議会[7]：薬系大学，製薬企業等の情報部門約130機関が会員で，伊藤四十二

教授(東大薬学部)の提唱により1955年に設立された．発足時より大学と製薬企業で構成された歴史を持つ．現在は永井恒司会長のもと，研究会や電子ジャーナルコンソーシアム事業，薬学会年会シンポジウムの開催[8]など活発な活動を行っている．

日本製薬情報協議会：大手製薬企業22社で構成され，2008年には情報科学技術協会の「優秀機関賞」を受賞した．本会の創設は1965年で，製薬文献データベースRingdocのユーザー会から始まり，データベース改善のためにトムソンに協力してきた．現在では情報全般の評価を行い，論文をホームページで公開している．[9] また，トムソンの製品についても品質評価に関与しており，副社長などの役員との会議を持ち，利用者としての要望や価格の維持を強く求めている．

このように，世界レベルでの巨大情報企業の戦略と研究現場での実際の情報ニーズを調整することは非常に難しい段階にきており，それらを行う情報担当者は薬学研究についての深い理解とともに，ビジネスの交渉力など多様な能力が必要である．そして，それらを行うために上記の2団体の果たす役割は大きい．

12 おわりに

日本の薬学研究は世界でも一流であるが，学術情報システムに関しては海外依存が多い．論文を投稿する学術雑誌も一流誌は海外中心であり，それを見るために高額な電子ジャーナル費用を払っている．そのため，学術雑誌が購入できなくなるシリアルズ・クライシス**という状況が進んでいるが，一向に解決の糸口は見えてこない．[10] このままでは，日本の研究自体への影響が懸念される．もはや個々の機関で対応できるレベルを超えており，国としての対策も望まれる．[11,12] また論文を探すためのデータベースも，JDream Ⅱを除いて海外中心であり，PubMedを例外として高額な契約費が必要とされるため，ライフサイエンス分野の主要5個を自由に使える環境にある研究者は少ない．このことは，PubMedを1〜2個のキーワードで検索して，すぐに論文が読める無料のオープンジャーナルから読んでいくというスタイルの普及も想像され，研究の質と効率化を大きく阻害する心配がある．さらに，医薬品の研究開発動向のデータベースもトムソンの独占に近い状況が進んでおり，将来的な高価格化が心配される．医薬品の研究開発情報は製薬企業にとって最重要課題であり，他社との差別化の必要性から海外のビッグファーマは独自にデータベースを構築して解析する傾向もある．この分野の市販データベースをどのように位置付け，どこまで投資を続けるのか，もしくは自社でも構築していくのか，将来的に判断が求められる．

参考文献

1) Kuhn T. S., "科学革命の構造," 中山 茂訳, みすず書房, 東京, 1971, p. 277.
2) 仲里猛留ほか, 情報の科学と技術, 60, 265-271(2010).
3) 小河邦雄, 薬学図書館, 51, 287-298(2006).
4) 松浦智佳子, 小河邦雄, 情報管理, 51, 408-417(2008).
5) 設楽真理子, 薬学図書館, 54, 138-142(2009).
6) 「インフォプロ」試験として情報科学技術協会の検定試験(1級, 2級)がある. http://www.infosta.or.jp/
7) 日本薬学図書館協議会 http://www.yakutokyo.jp/
8) 小河邦雄, 薬学図書館, 53, 226-231(2008).
9) 日本製薬情報協議会 http://piaj.sub.jp/ring/
10) 尾城孝一ほか, 情報管理, 53, 3-11(2010).
11) 日本学術会議, 提言 学術誌問題の解決に向けて 「包括的学術誌コンソーシアム」の創設, http://www.scj.go.jp/ja/member/iinkai/journal/index.html(2010).
12) 喬 暁東ほか, 情報管理, 53, 256-265(2010).

** シリアルズ・クライシスについての用語解説は，1142頁参照．