『蛋白質 核酸 酵素』バックナンバーの全文検索サービス公開にあたって

On the release of a full text search service of PNE back issues

川本祥子

大量の情報がデータや論文として生み出される生命科学分野において，総説やレビューは個々の情報を関連づけ，知識を体系化する重要な役割をはたしている．したがって，総説の全文がデータベースと組み合わせて検索可能になれば，その効果ははかりしれない．2008年6月より，統合データベースプロジェクトのサービスのひとつとして，『蛋白質 核酸 酵素』のバックナンバー全文を含むデータベース横断検索がはじまった．この国内初の試みの趣旨と概要について解説する．

Key words ■統合データベースプロジェクト ■文献検索 ■オープンアクセス

はじめに

21世紀に入り，人類をとりまく情報環境は大きく変革された．あらゆる情報が電子化され，インターネットにのって流通するようになり，その量は桁違いに大きくなってきたのである．米国IDC (International Data Corporation) の調査によると，2006年の世界のデジタルデータ量は161エクサバイト (エクサ＝10^{18}の単位) で，2011年にはその約10倍の1800エクサバイトに達すると予測されている (IDC白書「膨張するデジタル・ユニバース：世界の情報量に関する2011年までの成長予測」，**http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf**)．この数字は，20世紀までに人類が記した書籍の何百万倍ものデジタルデータが，たった1年で生み出されるということを意味する．情報爆発時代に突入し，われわれの生活はデジタルデータであふれるようになった．

生命科学分野も，時まさにこれと歩調を合わせるかのように，21世紀にさしかかるころからデータの大量生産時代に突入した．2003年に発表されたヒトゲノムは完全解読までに約20年を要したが，次世代シークエンサーの登場で，2010年には，ある個人の全ゲノムがわずか数日で解析できるようになるだろうといわれている．生命の設計図の完成は部品の機能解明に拍車をかけ，日々，膨大な量のデータが生み出されている．

このようななか，当然，研究者と情報をとりまく環境も変わってきた．そのなかでももっとも大きなものは，文献の検索や閲覧ではないだろうか[1]．1998年，生化学の代表的なジャーナルである *Journal of Biological Chemistry* (*JBC*) 誌が6カ月を経たバックナンバーの無料公開をはじめ，当時，たいへんな話題になったが，現在では多くのジャーナルが電子化された論文の全文を公開しており，電子ジャーナルという，冊子媒体をともなわないジャンルも出現した．米国National Center for Biotechnology Information (NCBI) のPubMed Centralや米国Stanford大学のHighWirePressなどを利用すれば，1000誌以上をいちどに検索することが可能である．研究者は実験室のかたわらにある情報端末で，世界中の論文を閲覧することができるようになったのである．

そして2008年，生命科学分野のジャーナルはさらに大きなターニングポイントを迎えることとなった．米国National Institutes of Health (NIH) が論文のパブリックアクセス方針を決定し，公的資金による研究論文は一定期間を経過したのちすべてオープンアクセスとすることを義務化したからである (**http://publicaccess.nih.gov/**)．これは出版業界に大きな波紋を投げかけたが，大勢はオープン化の方向にむかっており，ほかの国々でも順次導入されていく見込みである．わが国の生命科学分野の成果は，こうした話題の渦

Shoko Kawamoto
情報・システム研究機構 ライフサイエンス統合データベースセンター
E-mail：shoko@dbcls.rois.ac.jp
URL：http://lifesciencedb.jp

図1　統合データベースプロジェクトのポータルサイト“統合ホームページ”

「蛋白質 核酸 酵素」の検索には，全文検索とデータベース横断検索の2つのサービスがある（http://lifesciencedb.jp/）．

図2　“蛋白質 核酸 酵素 全文検索”のトップページ

中にある欧米のジャーナルへの投稿を中心に発展してきたが，国内をかえりみれば，本誌『蛋白質 核酸 酵素』をはじめとする日本語総説誌に幅広い領域から最新の研究が紹介され，生命科学というめまぐるしく発展する分野を理解するため多くの人々の助けとなってきた．現在は電子化も公開もなされていないこうした過去の文献がオンラインで利用できるようになれば，良質な知識の基盤となり，国内の研究にさらに寄与することが可能となるだろう．

統合データベースプロジェクト*1では，この生命科学分野全体の財産ともいうべき日本語総説誌を本棚に死蔵させ

*1　総合科学技術会議が掲げる戦略重点科学技術のひとつ“世界最高水準のライフサイエンス基盤整備”を実施する文部科学省のプロジェクト（2006～2010年）．大学共同利用機関法人 情報・システム研究機構のライフサイエンス統合データベースセンターが中核機関として，国内の生命科学に関するデータベース戦略の立案実施，データベースの統合化，ポータルサイト構築，基盤技術開発などを行なっている．

図3 “蛋白質 核酸 酵素 全文検索”での検索結果一覧の表示画面

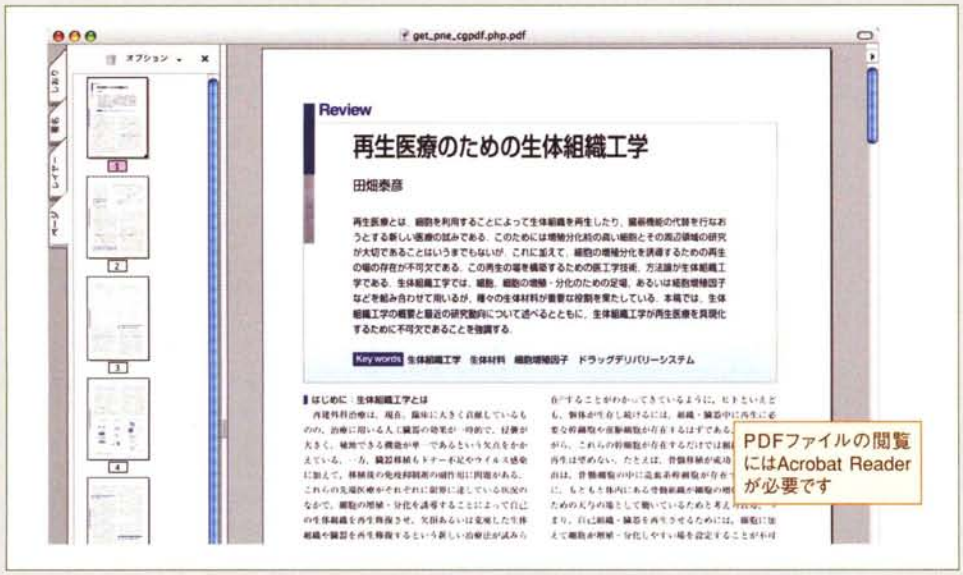

図4 “蛋白質 核酸 酵素 全文検索”のPDFファイル閲覧画面

ず有効に活用することはできないかと考え，『蛋白質 核酸 酵素』のバックナンバーの公開について，出版元である共立出版株式会社と協議を進めてきた．『蛋白質 核酸 酵素』は商業誌であり，公開によって販売数が減少するのではないかという危惧がないわけではない．しかし，激変する社会のなかで，過去の記事をオンラインで，しかも無償で公開することは，それ以上の価値を生み出すということで両者の意見は一致した．そこで，発行から約2年を経過した総説記事（Review，Short Review，解説，特集 など）についてPDFファイルを公開することが決定した．統合データベースプロジェクト側からは検索サービスなどの情報技術を提供し，約2年の準備期間を経て，ここに『蛋白質 核酸 酵素』バックナンバーの全文検索サービスを公開することができるようになった．記事の全文検索のほか，分子データ

図5 "生命科学データベース横断検索"の検索画面

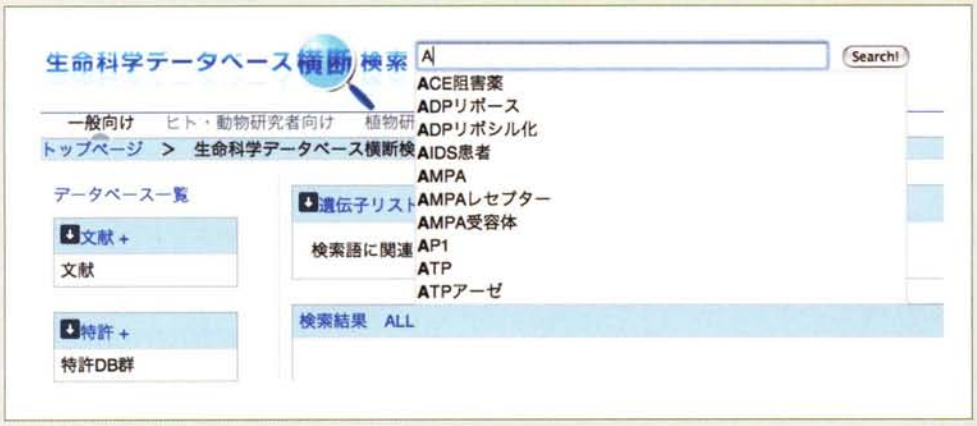

図6 キーワードサジェスト機能

ベースと組み合わせた横断的な検索も可能である．本稿では，以下，これらの利用法について解説する．

I『蛋白質 核酸 酵素』バックナンバーの全文検索

統合データベースプロジェクトでは，データベースのカタログやデータベース検索などさまざまなサービスを集め，ポータルサイト"統合ホームページ"(**http://lifesciencedb.jp/**，図1) から公開している[本シリーズ 2008年3月号，川本祥子・坊農秀雅の項 参照]．統合ホームページにアクセスすると，文献検索の項目に"蛋白質 核酸 酵素 全文検索"へのリンクがあり (図1)，ここをクリックすると"蛋白質 核酸 酵素 全文検索"のページ (図2) へとジャンプする．

それでは，実際に検索を行なってみよう（図3）．たとえば，検索キーワードとして“幹細胞 増殖因子”と入力する．Searchボタンをクリックすると，検索結果として，記事に“幹細胞”と“増殖因子”を含むもののタイトルがスコア順年代（降順）に表示される．このとき，検索結果にはタイトルの下に記事の一部が表示され，“幹細胞”と“増殖因子”が蛍光マーカーで示される．さらに，左側の発行年一覧には発行年別のヒット件数が表示され，検索結果を発行年別に絞り込むこともできる．全文検索なので，タイトルや要旨に含まれていないキーワードでも検索が可能である．ヒットしたもののなかから記事を選んでそのタイトルをクリックすると，PDFファイルが開き，記事の全文を読むことができる（図4）．なお，PDFファイルを閲覧するにはAcrobat Reader（無償）が必要となる．また，“蛋白質 核酸 酵素 全文検索”では，現在は2005年末までに発行された例月号の総説記事について全文を検索し閲覧することができるが，コラムやシリーズ，また，増刊号の記事については検索できない．

ここで，利用にあたっての注意点を示しておきたい．この検索サービスは出版社がもつ出版権の範囲で行なわれており，著作権法にもとづき，その利用は個人での利用に限定されている．よって，PDFファイルのダウンロード，ならびに，印刷はできないようになっている．もちろん，コピーや改変，転載なども禁止されている．もし違法な使用

表1　全文検索が可能なWebサイトおよびデータベース

データベース名	対象レコード数	バージョン	更新日（収載年）
『蛋白質 核酸 酵素』	1,577		1985～2005年
文部科学省特定領域研究「ゲノム」研究報告書	173		
特許・実用新案広報	369,647		2004～2007年
英語ウィキペディア	327,175	−	−
日本語ウィキペディア	31,965	−	−
KEGG	45,628	44	2007年10月 1日
PDBj	45,379	v3	2007年 7月31日
RefSeq	286,797	release20	2006年11月21日
H-Inv	187,156	release5.0	2007年12月26日
JSNP	76,067	release31	2007年 6月25日
DBTSS	31,626	release6.0	2007年 9月15日
HUGE	2,038		2007年 3月30日
NEDO	444		2004年 3月12日
BodyMap	85,769		
FANTOM	33,424	ver3.0	2005年 2月21日
ROUGE	2,169		2006年 3月30日
RedClover	38,271		2007年 1月11日
RAP DB	51,470	release2	2007年11月22日
RPSD	366		2005年 2月10日
CyanoBase	13,445		2007年11月14日
RhizoBase	41,654		2008年 1月15日
GTOP	49,821		2006年12月21日
Mycoplasma penetrans genome	1,073		2006年 3月14日
JAPIC	26,692		2007年 9月 1日
GGDB	195		
Lipid Bank	5,952		2007年 6月19日
Entrez Gene	103,473		2008年 3月
PIR	384,953	3.46	2008年 2月26日
OMIM	19,570		2008年 5月16日
PubMed（外部情報）	−	−	−
米国特許（外部情報）	−	−	−
欧州特許（外部情報）	−	−	−

が多発すれようであれば，このサービスの継続は困難になることを理解してほしい．

II 生命科学データベース横断検索

『蛋白質 核酸 酵素』のバックナンバーを検索できるもうひとつのサービスが，生命科学データベース横断検索である（図5）．統合ホームページのDB検索の項目に“生命科学データベース横断検索”へのリンクがある（図1），2008年6月に公開を開始したこのサービスは，複数のデータベースを一括して検索できる機能をもった，データベース統合化における重要な柱のひとつである[2]．総説や報告書，特許などの日本語の文献情報と分子データベースとをあわせて検索することを可能にした国内初の試みであり，これまで並べて見ることのできなかった情報を一括して取得することで，従来にない価値を生み出すことに成功した．そのなかで，『蛋白質 核酸 酵素』のバックナンバーは分子データベースどうしを結びつける重要な役割をはたしている．

生命科学データベース横断検索における代表的な機能を紹介する（図5）．ここでは，検索キーワードとして“転写因子”と入力した場合を示している．

機能1：検索結果を利用者の用途に応じて並び替えることができる．たとえば，“植物研究者向け”をクリックすると，植物のデータベースが上位にくるようになる．文献を中心に検索を行なう場合は“一般向け”とする．

機能2：検索結果を画面の左側にあるデータベースごとに絞り込むことができる．

機能3：検索オプションでは，遺伝子IDや著者名を限定した検索ができる．

機能4：キーワードサジェスト機能をもたせることで，専門分野以外の情報の入手も容易になっている（図6）．ここでは，“A”と入力することで，Aではじまるキーワードがポップダウンリストで示されている．このなかから，自分の調べたいキーワードを選ぶことができる．

機能5：検索キーワードに日本語を使うことができ，このとき，日英変換が自動に行なわれる．図5の例では，“transcription factor”も検索キーワードとなっている．なお，このとき変換辞書としては，おもにライフサイエンス辞書を利用している．

この生命科学データベース横断検索は，基本的にはGoogleライクな検索であるが，検索エンジンにはオープンソースのHyper Estraierを利用しており，公開時点での検索対象は国内外あわせて32である（表1）．今後は，さらに収録するデータベースを拡大していくほか，文献（具体的には，雑誌のバックナンバーなど電子化されていない文書や，公開されていても自由に検索ができない文書，散逸している文書）へのアクセスの確保をより一層強めていく予定である．また，現状はキーワードのヒット数に頼っている検索結果であるが，検索精度をさらに向上させるための情報技術の開発や，辞書やオントロジーの構築，索引づけは，短期間に達成すべき課題である．できるだけ多くの方に使ってもらい，意見を聞かせていただきたい．

おわりに

電子化された情報が爆発的に増加するなかで，必要な情報を見つけ出すことはさらに難易度が高まっているとさえ感じられる．その原因には，情報量の多さだけでなく，情報はさまざまな形をしていて，1カ所に留まっていないなどの性質もあわさっている．Googleをしのぐ検索は，今後も容易には実現されないかもしれない．統合データベースプロジェクトでは，利用者が情報の取得に払ったコスト以上に得られた効果のほうが大きいと感じられるような検索サービスの開発をめざすと同時に，文献やデータベースの公開を促進していくことで，生命科学分野での情報格差を減らすことに努力していきたいと考えている．

最後になりましたが，今回のバックナンバー公開の趣旨に賛同し実現にご協力いただいた共立出版株式会社に感謝いたします．

【文献】
1) 大久保公策: 蛋白質 核酸 酵素, 52, 1027-1031 (2007)
2) 高木利久: 蛋白質 核酸 酵素, 52, 1388-1393 (2007)

川本祥子
略歴：1995年 大阪大学大学院理学研究科博士後期課程 修了，同年 大阪大学細胞生体工学センター 助手，2002年 九州大学生体防御医学研究所 助手，2003年 国立情報学研究所 プロジェクト研究員，2004年 奈良先端科学技術大学院大学 寄付講座教員，2006年 情報・システム研究機構新領域融合研究センター 特任准教授を経て，2008年より情報・システム研究機構ライフサイエンス統合データベースセンター 特任准教授．

本シリーズは今回をもって連載終了となります．1年間，ご愛読をありがとうございました．ご意見・ご感想などをE-mail：pne@kyoritsu-pub.co.jpまでお寄せいただければ幸いです．
（編集部）