有機化合物のスペクトルデータベースの開発と公開サービス

―　大規模データベースの運用の継続と成功の秘訣　―

齋藤 剛*、衣笠 晋一

産業技術総合研究所の有機化合物スペクトルデータベース（SDBS）は1982年に開発を開始し、以来30年間変わらない部分と大きな変化を遂げた部分を混在させつつ高度化されてきた。標準スペクトルとして信頼性の高いものを収録すること、1種類の化合物に複数種類のスペクトルを収録することの二つの基本コンセプトと、汎用化合物を対象とする点は、開発当初から現在まで変わらず引き継がれている。一方、データベースを収集するプラットフォームと公開形式は大きく変わった。データのウェブ公開に伴ってユーザーからの声を取り上げ、各種の依頼や指摘に対応するようになったことも、大きく変わった点である。長期間にわたって開発と公開サービスを継続し、現在ウェブを通して多くの研究者、技術者、教育者、学生らによって利用されるにいたった。データベースの全体構想から、構造の設定、データの収集方法、データの公開方法等主要なプロセスを統合的、構成的に記述する。

キーワード：スペクトル、データベース、核磁気共鳴、赤外分光、質量分析、化合物情報、ウェブ

Development and release of a spectral database for organic compounds

– Key to the continual services and success of a large-scale database –

Takeshi Saito* and Shinich Kinugasa

The research activities of spectral database for organic compounds (SDBS) in AIST started in 1982. Since then, many parts of research activities have changed while the other parts have remained unchanged for almost 30 years. The unchanged parts since the start of this project are the two principles that the spectral data with high authenticity should be compiled in the database as the standard data and that several kinds of different spectra should be compiled for each compound, and the concept that compounds used commonly in industries and societies are object of compilation. On the other hand, the computer system used for database management and the ways for data release have changed completely over time. After the data have come to be opened to the public through the Internet, we have started to take considerations of comments, requests and indications from users. SDBS has had innumerable Internet accesses from many researchers, engineers, educators and students from all over the world. In this paper, the total framework, the structure of the database, the method for its data compilation and the ways to release the data to the public are described with analysis and clues of long time continuance and success of SDBS activities.

Keywords：Spectrum, database, nuclear magnetic resonance, infrared, mass, chemical information, web

1　はじめに

化学物質の信頼性の高い分析が、産業界をはじめ社会のいろいろな場面で要求されている。とりわけ化学分析の中で核磁気共鳴スペクトル、赤外分光スペクトル、あるいは質量スペクトル等は有機化合物を同定するための有力で基本的な情報である。新規化合物の開発や材料中の未知化合物の分析等化合物を同定する必要がある現場で多くのスペクトルの測定と解析が行われている。一般に、測定して得たスペクトルデータを標準スペクトルデータと照合することによる同定は、最も信頼性が高い分析方法の一つである。この方法は広く用いられており、このような標準データおよびそのデータベースの果たす役割は大きい。

産業技術総合研究所（産総研）における有機化合物のスペクトルデータベース（Spectral DataBase System for organic compounds, SDBS）の開発は1982年に旧工業技術院のプロジェクトとして、①標準スペクトルデータとして信頼性の高いデータを収録すること、② 1種類の化合物に対して複数種類のスペクトルデータを収録すること、の二つを基本コンセプトとして開始された。最大6種類のスペクトル、すなわちMS（質量スペクトル）、^{13}C NMR（炭素13核磁気共鳴スペクトル）、^{1}H NMR（水素核磁気共鳴スペクトル）、IR（赤外分光スペクトル）、Raman（ラマン分光スペクトル）とESR（電子スピン共鳴スペクトル）のデータを研究所自らが取得することと、その化合物の情報管理をすることが基本的な内容である[1]。研究活動を開始してから30年近くが経過する中でRamanとESRはデータの収集を中止し、現在では、MS、IR、^{1}H NMRと^{13}C NMRの4種のスペクトルの収集を継続し、あわせて化合物情報管理と公開の活動を行っている[2]。

産業技術総合研究所　計測標準研究部門　〒305-8563 つくば市梅園 1-1-1 中央第3
National Metrology Institute of Japan, AIST　Tsukuba Central 3, 1-1-1 Umezono, Tsukuba 305-8563, Japan　* E-mail: takeshi.saito@aist.go.jp

Original manuscript received August 10, 2010, Revisions received October 12, 2010, Accepted November 2, 2010

1997年に旧工業技術院のプロジェクト[3]により、インターネットを通したウェブでのデータ公開[4]を開始した。2010年4月現在で公開しているデータは、化合物数の総数が33,000あまり、それぞれのスペクトルの数と割合は図1に示すとおりであるが、スペクトルの総数は約10万である。現在の有機化合物スペクトルデータベースはウェブでアクセスしてくるユーザーが主たる利用者である。データ公開を開始して以来、インターネットで多くのアクセスを得ており、過去3年間は1日の平均ページビューが10万件を越え、産総研が公開している「研究情報公開データベース（RIO-DB）」の中で、アクセス数が群を抜いて高く、2009年度末に公開以来延べ3億ページビューを記録した。年度別アクセス数の推移を図2に、公開しているスペクトル数の推移を図3に、それぞれ示した。社会におけるインターネットの利用の拡大と、このデータベースの認知度の向上により、アクセス数はこの10年間毎年著しい伸びを示している。これらのスペクトルデータを、教科書[5]、参考書[6]、テスト問題等へ利用したいという要望は絶えず届き、データベースの中にある誤りを指摘してくれるユーザーもいる。

このデータベースの開発のシナリオを図4に示す。ここには、データベースを構成する種々の要素を列挙し、それらがデータベースの主要な特性である基本構造、網羅性、信頼性、利便性とどのような関係にあるかを示した。あわせてデータベースの運用に当たって重要となる要素も示した。それらの要素をどのように統合したか、そのプロセスを以降の章で記述する。

2　データベースの構造

2.1　データベースの基本構造の重要性

このデータベースは一つの化合物に対して複数の種類のスペクトルデータが閲覧できる構造をとった。これを実現するために、図5に示すように化合物情報と6種類のスペクトル情報を収録したデータベースを独立に作成し、化合物情報データベースを中心にして全体を統合した。

この作業を円滑に行うために、データベース構築の現場ではいくつかの種類の管理番号を準備した。すなわち、各

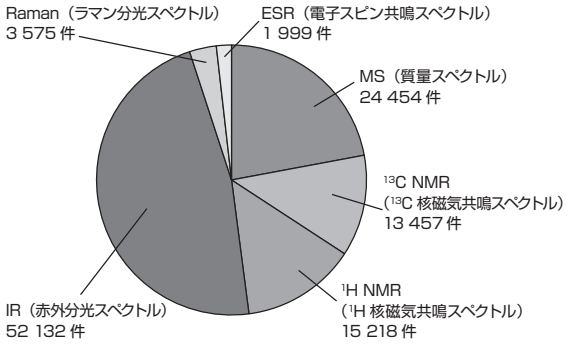

図1　2010年4月現在で産総研有機化合物スペクトルデータベース（SDBS）においてウェブ公開しているスペクトルの割合と公開数

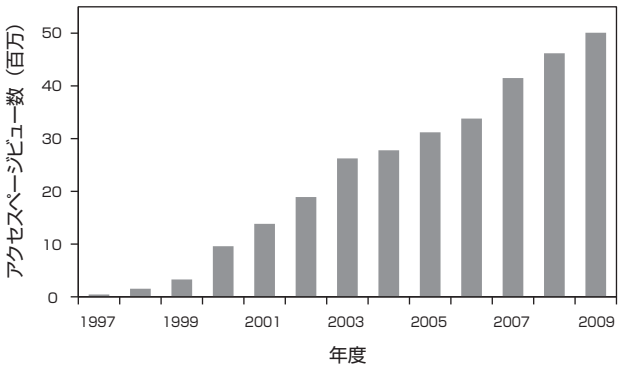

図2　産総研の有機化合物スペクトルデータベース（SDBS）のウェブ公開以来の年度別アクセスページビュー数の推移

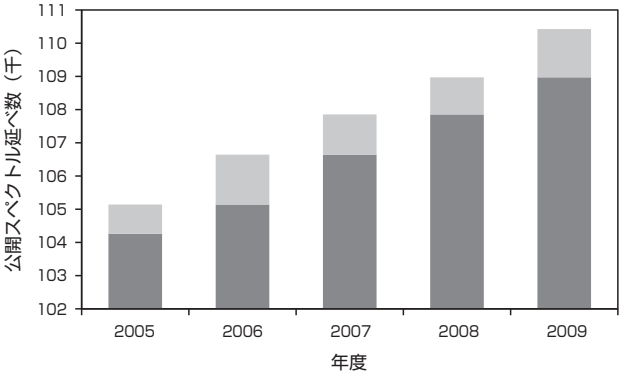

図3　産総研の有機化合物スペクトルデータベース（SDBS）の過去5年間のスペクトル数の推移
各年度の新規公開分は薄い色で示した。

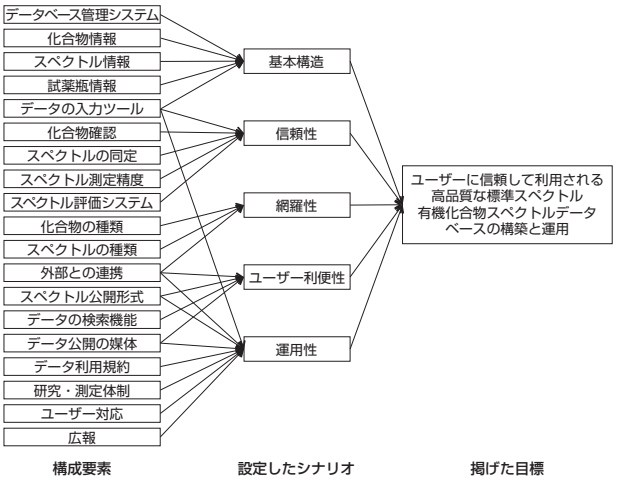

図4　有機化合物スペクトルデータベース（SDBS）の構築と公開のためのシナリオ

試薬瓶に対して与えられるビン番号、取得した各スペクトルに対して与えられるスペクトル管理番号と、そのうちデータベースに登録されたものに与えられるスペクトルコード、各化合物に対して与えられる SDBS 化合物番号（このデータベース内では単に SDBS 番号と呼ばれている）である。スペクトルコードは、MS、IR、^{13}C NMR と ^{1}H NMR で別々に管理しており、それぞれの化合物情報やスペクトルデータは独立したデータベースとして切り出すことも可能である。これらの番号に対応した化合物情報やスペクトル情報はリレーショナルデータベースの形態で管理しており、現場での作業が円滑に行えるようにした。

特に化合物に固有な SDBS 化合物番号を採用したことは、このデータベースの特徴である。この番号は単なる化合物の管理番号ではなく、この採用によって、データベースの中で化合物辞書を一つのデータベースとしてスペクトルデータベースから独立させたので、柔軟な変更が可能となり、現在でも有用な化合物情報としてこのデータベースの運営を可能としている。この画期的な SDBS 化合物番号を採用した背景には、旧工業技術院時代に先行的な研究を行っていたガスクロマトグラフィーデータ委員会、赤外データ委員会、NMR データ小委員会等でのさまざまな成果があり、またこのデータベースの基本構造の設計時に多くの知見の蓄積があり、そのことが 30 年近くたった現在でも機能するデータベースを可能としたと考えられる。

このデータベースではすべて産総研が入手した化合物に対して実際にスペクトル測定をしてデータを取得することを原則とした。入手した化合物の試薬瓶ごとに固有の番号であるビン番号を付与した。一方、SDBS 化合物番号は化合物に対する固有の番号であることから、ある新たな試薬瓶に入っている化合物がすでに登録された化合物と同じ化合物であるかどうかの判断を行う必要があり、既登録の化合物と一致しなかった場合にのみ新しい SDBS 化合物番号を付与した。化合物数が多くなかった初期の時点ではこの作業はさほど困難なものではなかったと思われるが、これらの確認作業には疑問や問題が生じることもしばしばあった。3 万件に及ぶ化合物を対象とし、しかも開発当初とは異なり化合物の名称やその構造がより複雑になった現在においては、化合物に対して固有の番号を付与することはより労力と時間を要する作業となり、専門化学者がこれらを行う必要性があるために、スペクトル収集そのものよりも労力と時間を費やすようになっていった。この問題を解決し、専門家のリソースをスペクトル測定、収集、評価により多く確保するために、SDBS 化合物番号の確認方法として、多くの化合物情報項目の一致検索を行った結果、データベースにすでに登録されている可能性があるという判定が示されたときにのみ、複数の専門化学者が疑問点を重点的に確認する方式を確立した。これにより、作業の質を落とすことなく、SDBS 化合物番号を付与するために専属の専門家が不要となっただけでなく、データベースの化合物登録作業が円滑になり、化合物番号の重複が起きにくくなった。

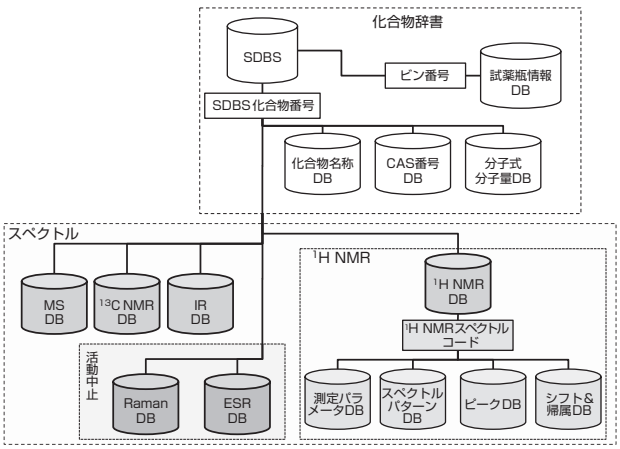

図 5　有機化合物スペクトルデータベース（SDBS）の構造
図中の SDBS には、化合物番号である SDBS 化合物番号と、その化合物の元素数等が記録されている。それぞれのスペクトルデータベースを代表して ^{1}H NMR データベースの構造を示した。すべての情報は、SDBS 化合物番号を利用したリレーショナルデータベースで関係づけられている。

2.2　データベースの運用とプラットフォーム変更に際しての判断

開発当初の 1980 年代はこのデータベースの運用を大型コンピュータで開始した。この時期が日本の Windows PC の源流である NEC PC-9800 シリーズの発売開始時期と変わらないことからみても、大型コンピュータで運用を開始したことは当然であろう。しかし、この大型コンピュータ（FACOM MSP）は旧工業技術院の方針により運用が 1999 年 3 月末で終了したため、他の大型コンピュータへ移行するか、あるいはパソコンへ移行してデータベースを継続するか、あるいは活動を終了するか選択を迫られた。この時点で私達はデータの管理を行うシステムに Windows PC を採用することに決定し、新たにパソコンを利用したデータ入力ツールを開発することにより活動を継続した [7]。多くのソフトウエアで MS-DOS から Windows への移行がうまくいかずに再構築しなければならない等の困難に遭遇したが、このデータベースを全く異なる環境のプラットフォームに移すことに成功した。このときに大型コンピュータにこだわっていたら、現在のデータベースのようなデータの入力や管理を助ける様々なツールを利用することに支障を来したであろう。データ収集を Windows PC で行うようにしたことでデータの管理が格段に容易になった。

3　収集するデータの選択

3.1　化合物の選択の戦略

このデータベースは、広い分野で化合物同定等の分析に利用することを念頭に構築したものであるため、多くの人が頻繁に使用する市販試薬を中心にスペクトルデータを収集することとした。ウェブで公開しているデータ数は図1に示したが、公開にいたらなかったスペクトルデータも数多くあり、測定に供した化合物は試薬瓶でのべ39,000本以上になる。

このうち、10,000本以上の試薬は東京化成工業株式会社から無償で提供されたものであり、収集した試薬数のうちでは､同社から提供されたものが最も多い。したがって、試薬選定は同社の新規試薬開発方針に沿った部分があるが、それは間接的にユーザーの現行のニーズを反映したものといえる。研究開発現場で化学合成等を行って種々の化合物を調製する場合も、その原料は市販試薬であることが多く、化学工業における基盤的な試薬を同社から多く入手することができたことは貴重な支援であった。

2001年以降はこれとは別に農薬や劇物を中心にした化合物のスペクトル収集を開始した。すなわち法規制があるもの等危険物のスペクトル情報を多く収集し発信することは公的研究機関の重要な役割であり、最近まで徐々にその数を増やしている。食の安全への関心も高まっており、その後も農薬等重点的に収集する化合物を選択する戦略は重要と考えている。

3.2　データ形態（デジタルデータ）選択の先見性

データ収集の形態にかかわる最も重要な選択が開発当初に行われた。それは今となっては当然のことであるが、このデータベースはすべての情報が座標データとしてコンピュータ上でデジタル化して収録されたことである。1970年代にはスペクトルデータ集として冊子体のデータ集が活用されていた。測定データはデジタル化により取り扱いが容易になることは認識されていたが、コンピュータの容量の制限等からデジタル化によって情報の一部が失われる等の問題もあり、紙媒体等に記録するアナログデータの取り扱いが主流であった[1]。データのデジタル化には、NMRの測定を例にとると、その当時でも一つのスペクトル当たり数万のデータポイントで構成されており、30年前のコンピュータのディスクやメモリ容量等の条件を考えると、デジタルデータでのデータベース化は大きな決断だったはずである。当時の工業技術院の大型コンピュータがなければ実現がむずかしかったと考えられる。このような条件では、個々のスペクトルの測定だけでなく、それをデータベース化するためには多くの困難があり、収録するデータ量を最小限に抑える工夫を合わせて行うことで実現できた。実際に、^{1}H NMRのスペクトルパターンの座標データをデジタル化して収録したのは世界初[8]であり、取得したデータの必要な部分のみを切り出して収録する方法を取ることでデータ容量を圧縮した。^{13}C NMRの測定データは、ピーク値を規格化した強度と半値幅の値を収録し、ピーク形状をローレンツ関数で仮定してシミュレートして表示した。IRとRamanは得られたスペクトルの各点の座標データを収録し、MSは質量数とその強度を収録した。ESRもスペクトルをデジタル化したが、論文のデータをカーブリーダで読み取って、座標をデジタル化したものもあった。^{1}H NMRには、化学シフトとスピン結合定数を利用したスペクトルシミュレーション機能を備えた[9]。産総研になってからは、NMRは^{13}C NMRも^{1}H NMRも不純物由来のピークやノイズを含めたすべてのデジタルデータをスペクトルデータとして収録したことで、ユーザーはピークの信号強度やその時のノイズレベルまで確認することができるようになった。1997年に旧工業技術院からウェブ公開を行ったことはすでに述べたが、開発当初にもしもスペクトルのデジタルデータを収録していなければ、予想されるスペクトルデータのデジタル化に対して、アナログ収録した化合物に対しては再測定せざるを得ない状況になったと思われる。

3.3　質と量のバランスにおける高品質へのこだわり

このデータベースのスペクトルデータは、ESRと^{1}H NMRの一部に論文の情報から作成したスペクトル情報があるのを除き、開発当初からすべて当所で測定、評価したデータを収集する方式をとった。すべてのスペクトルデータに対して、品質に責任を持って公開していくためにはこの方式が最も信頼性が高い方法である。この方式は、公開するデータの品質の確実性に利点がある一方、公開できるデータ量は限られてしまう。多くのスペクトルデータを公開すること、すなわち網羅性を高めることはデータベースの重要な要素の一つである。この質と量という異なる二つの価値をどのように調和させていくか、また、データベースをこれら二つの価値軸のどの位置に設定するか、データベースの存在意義にもかかわる大きな問題である。このデータベースではまず標準データとして一定の質を確保し、その上で時間をかけて量的な要求に応えるという方針をとった。

データベース構築に当たってスペクトルデータの信頼性確保のために評価基準を定めたが、その例を以下に示す。^{1}H NMRではテトラメチルシラン（TMS）を化学シフトの基準として利用するだけでなく、スペクトルの分解能の判断基準にも利用した。TMSのピークが尖鋭化していれば、化合物のピークの分解能が悪く見えても、それは測定の不備のためではなく、その試料が示す特性であると判断できる。IRスペクトルの場合には干渉ノイズが無いことや水ピー

クが無いこと、あるいはベースラインが大きくうねらないことをスペクトル評価の基準とした。このような基準を、それぞれのスペクトル測定を担当した研究者が独自に設定してきた。

3.4　データ登録の方針

同じ条件で測定されたスペクトルは、最も信頼性が高いと評価されたもののみ公開した。MS スペクトルは直接導入法を採用したので、化合物に対する測定条件が一つに決まるため、最も良いと判断されたスペクトルを化合物に対して一つだけ登録した。IR は、例えば固体試料の KBr 錠剤法とヌジョール法等、同じ化合物でも異なる条件でスペクトルを測定したため、それぞれの条件で最も信頼性の高いものを登録した。^{13}C NMR スペクトルは ^{1}H 核とのスピン結合がなくなる条件で測定したため、一つの化合物に対してスペクトルを一つ登録した。

^{1}H NMR は同じ試料と溶媒の組み合わせでも、得られるスペクトルのパターンは測定する周波数に依存する。活動当初からスペクトル収集に最も利用された測定共鳴周波数は 90 MHz であるが、この周波数でのスペクトルが複雑で解析が困難な化合物については、より単純なスペクトルが得られる 400 MHz で測定した。実際のデータから抽出した化学シフトとスピン結合定数を利用して、異なる周波数でシミュレートしたスペクトルを示すことも重要であったことから、スペクトルとは独立に化学シフトとスピン結合定数を収集し、シミュレートスペクトルを示せるようにした。

NMR はスペクトル自身に加えて、それぞれのスペクトルを解析して帰属情報を付与した。特に ^{1}H NMR は測定する共鳴周波数によってスペクトルのパターンが変わることから、普遍的な値である化学シフトと帰属情報を付けることが不可欠であった。これらの情報がなくては、異なる共鳴周波数で測定したスペクトルとの比較が困難であり、化学シフトと帰属情報は ^{1}H NMR のスペクトルデータベースとして最も価値の高い情報といえる。

測定に供した化合物に関する情報は可能な限り多くを収集し登録した。化合物の構造が複雑であればあるほど、ある一つの化合物に対して異なる複数の名称が使われるのが普通である。データベースの利用者がどの名称で検索しても対応できるように、化合物自身に関する情報は網羅的に登録した。

3.5　収集するスペクトルの種類

スペクトルの収集は開発当初 6 種類について行い、現在は 4 種類について継続している（図 5）。開発当初の 1980 年代に汎用的に利用されていた分析法は他にもあったはずだが、例えば紫外可視分光法等のスペクトルデータはデータベースに採用しなかった。採用した 6 種類のスペクトルは当時の研究所の測定装置や研究者に依存した選択であったと考えられる。その後 Raman と ESR のデータ収集を継続しなくなった原因は、担当する研究者や稼働できる装置の問題も一部あるが、潜在的な需要が活動中止の大きな要因であったと推察する。ところが現在は学術的にも産業的にも Raman スペクトルデータの需要が拡大しており、その点でこのデータベースは十分な対応ができていない状況にある。一方、MS、^{13}C NMR、^{1}H NMR と IR は 1980 年代から現在に至るまで大きな需要がある。この点は、特にこのデータベースがウェブで公開されたあとはユーザーからのアクセス数によって直接的に需要が確認されている(図 2)。

4　データ公開の方針

4.1　ウェブによるデータの公開

1997 年に産総研のウェブサイトから MS、^{13}C NMR と ^{1}H NMR のスペクトルデータの公開を行い、1 年遅れて IR と ESR を公開した [10]。現在は Raman を含めて 6 種類のスペクトルデータを公開している。このデータベースのウェブ公開を開始したときは NCSA Mosaic や Netscape Navigator のようなウェブブラウザが開発され、多くの人がウェブを利用するようになっていたものの、ウェブ回線が現在に比べてまだ整備されておらず、またブラウザの表示能力も不十分だった。ウェブでデータを公開するに当たっては、スペクトルをより効率的に表示することが重要な課題であった。このことからスペクトルや化合物の構造を表示するために最も小さいサイズの画像表示形式であり、インターネット回線に負担の少ない GIF ファイル形式を選択した。日本国内でのウェブアクセスは高速化しているが、必ずしも高速化に対応できていない世界のユーザーを考慮し、現在もまだこの形式を継承している。

画像表示形式である GIF 形式でデータを公開するもう一つの理由はデータの保護にあった。すなわち、このデータベースの知的財産である座標データを不正に取られてしまうことを防ぐ対策であった。座標データからは、高品質のスペクトルの再構成等が容易に可能であるが、画像表示形式からは、それを超す品質の情報を作ることができない。これまでに短時間に系統的かつ網羅的にデータを取得することを目的としたアクセスが数回あったことが分かっているので、データを公開するために取った保護策は有益であった。今後は、このような不正アクセスへの対策を講じた上で、座標データを利用したウェブ上でのスペクトル拡大機能等を装備することも可能となろう。

ウェブサイトからのデータ公開にあたって、公開ページの言語情報は英語表記とした。これは、収集した化合物名称が英語であり、その他の情報は言語に依存しなかったこ

とから可能であった。現在は、国内ユーザーへのサービスとして、アクセスするコンピュータの言語設定が日本語の場合には、フレームに日本語で説明が表示される仕組みにした。

重要な情報であってもこのデータベースで整備し切れないものに対しては、積極的に他の利用可能なデータベースとリンクを張ってユーザーへの便宜を図った。2006 年からは東京化成工業株式会社のオンラインカタログとこのデータベースのリンクならびに、科学技術振興機構の化学物質リンクセンター [11] とこのデータベース間のリンクを張った [12]。日本語を利用した化合物検索や構造式検索等、現在このデータベースで独自に整備し切れていない部分を補完するようにした。

ウェブでデータ公開したメリットの一つとして、公開データを一括管理できることが挙げられる。公開用サーバのデータを更新することで、すべてのユーザーに対して同時に同等のサービスを提供することが可能となった。ウェブを介してユーザーからのコメントが直接開発者に寄せられることも、ほかの研究と異なる特徴である。

ウェブ公開以前は、1989 年からオンライン [13] で、1991 年からは CD-ROM 媒体にデータと検索プログラムを入れたデータベースソフトウエア [14] として販売した。この CD-ROM 媒体を利用できたのは国内数十件程度の特定ユーザーのみであった。この形態では、提供されるデータはある時点までに収集されたものに限られ、長期間にわたって保存できるが、データの更新やソフトウエア改修等への対応は難しい。また、CD-ROM では所有する一部のユーザーに対してのみのサービスに限定される。しかし、MS-DOS で動く CD-ROM で検索、表示可能にしたことで、現在のウェブ公開に必要な要件をあらかじめ検討できた。このことは、開発前に先導的研究を行ったことによりこのデータベースを適切に設計できたことと類似している。CD-ROM での公開は、このデータベースをウェブで公開するための重要な先導研究であったと考えられる。

4.2　ユーザーの解析と公共財としての役割

このデータベースのユーザーを解析するために、2009 年度のアクセスログを国別認識記号で集計解析した結果を図 6 に示した。延べ 5 千万件超のページビューのうち国内からのアクセスは約 14 % であり、一方最もアクセスが多かったのは北米地域であった。商用を表す「.com」、ネットワークを表す「.net」等、特定地域に限定されないドメインは、国別識別コードとは別に集計した。日本国内からのアクセスに関してアクセス元のドメインを集計したところ、図 7 に示したように大学等の学術機関である「.ac.jp」からのアクセスが最も多く、ネットサービスやインターネットサービスプロバイダ等に発行される「.ne.jp」、一般企業のための「.co.jp」がそれに続いた。これらの中で学術機関とネットワークプロバイダー等のドメインからのアクセスは、一年の中でも 3 月と 8 月は、最もアクセスの多かった 6 月の半分以下と、季節変動が激しかった。一方、一般企業からのアクセスは、多少の変動はあるものの一年を通しておおむね同じ程度のアクセス量があった。このことから、学生の夏休みと学期末と重なる時期にアクセス数が低下する傾向にあることがわかる。ネットワークプロバイダー経由のユーザーのアクセス傾向が、学術機関からのユーザーのアクセス傾向と類似していることから、ネットワークプロバイダー経由のユーザーも多くは学生であることが示唆され、全体として多くの学生に利用されていることがわかった。

このようにこのデータベースは多様なユーザーに使用され、産総研のような公的研究機関が提供する公共財の役

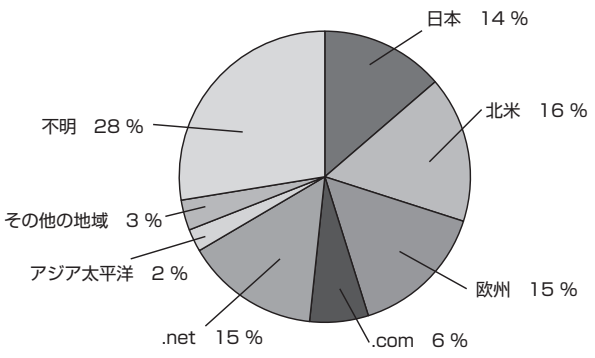

図 6　2009 年度 1 年間にデータベースへアクセスしたユーザーの地域ドメインの割合
地域を特定しないドメインのうち、「.com」と「.net」は独立に集計した。不明は、アクセス元に IP アドレスを利用する等、ドメイン名が特定できなかったアクセス元である。

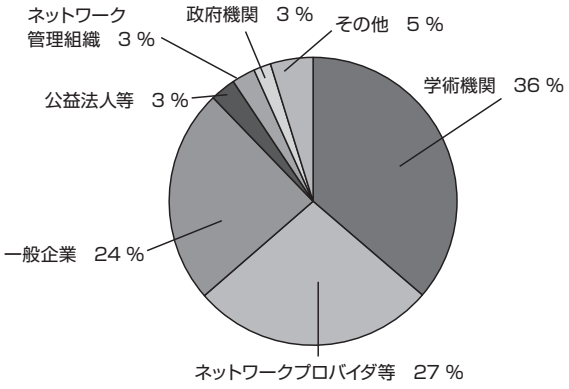

図 7　2009 年度 1 年間にデータベースに国内からアクセスしたユーザーのドメインごとの集計
学術機関、ネットワークプロバイダー等、一般企業、公益法人等、ネットワーク管理組織、政府機関は、それぞれ「.ac.jp」、「.ne.jp」、「.co.jp」、「.or.jp」、「.ad.jp」、「.go.jp」ドメインに対応しており、これ以外の「.jp」ドメインからのアクセスはそのほかにまとめた。

割を果たしている。一般にデータベースは多くの情報を蓄え、その中から必要な情報を効率的に検索することでその力を発揮する。このようなデータベースの構築と維持には、多くの資源と時間が必要となり、このデータベースも例外ではない。しかし、データベースの利用者に対して、この開発維持の対価を要求すると、利用することができるユーザーは限定されてしまう。このデータベースは、公共財として無料で公開することで、産業界はもとより、スペクトルデータの利用法を学習する人も含む多くの人々に機会を与える役割を担っている。無料で公開することで、数多くの企業がデータベース構築に付随するコストをかけずにスペクトル情報を化学分析の現場で利用することができ、これによって化学分析のコストが低減される。すなわち、このデータベースは、産業を支える知的基盤としての役割を果たしている。さらに、実際に大学からの利用が国内では全体の 36 %あることや、教科書や研修資料にスペクトルを利用したい旨の依頼が数多く寄せられるように、国内外でスペクトル利用法の教育にも広く使われており、社会全体に大きく貢献していることがわかる。

4.3 ユーザーからのコメントへの対応

データベースを公開したことによりユーザーからは多くのコメントがメールで寄せられる。研究の多くは論文の形で世に問われて評価されるが、このデータベースは世界中にいるウェブユーザーに直接問われており、ユーザーから得られる評価の表れの一つがコメントである。ユーザーからのコメントを真摯に受け止め、それを活用していくことはデータベースの今後の方向性を見極め、さらに発展させるために重要だと考えている。

コメントは利用許諾の申請、そして技術的な質問や指摘に分類される。感謝のメールも多く届き、私達もそれらを見ると勇気付けられる。

技術的な指摘は NMR の帰属の間違い、測定条件の問題等がある。技術的な間違いの指摘を受けたときには、すぐにそのデータの精査をする。この段階でユーザーからの指摘に対する判断ができない場合には、スペクトルを再測定することもある。スペクトルデータを再精査した結果、ユーザーのコメントが正しいと考えられる場合には、すぐに修正を行う。ウェブで公開しているデータベースの情報が正しいと判断した場合には、指摘してくれたユーザーに対して説明を行った上でその情報の公開を継続する。再測定には同じ化合物を可能な限り入手して対応しているが、それができない場合にはそのデータを取り下げることもある。

ユーザーからの、ウェブで公開しているスペクトルの画像、すなわち GIF ファイルの他の資料への利用許諾に関する申請には可能な限り利用してもらえるように対応している。いずれのコメントも、このデータベースで公開しているデータの質が確保されていることの一つの現れと考えており、今後ともこのようなコメントに対して迅速に対応できる体制を維持していくことが重要と考える。

アクセスログの解析から教育機関からのアクセスが多いことを示したが、教科書に頻繁に現れる化合物のスペクトル、特に ^{1}H NMR に関する問い合わせが届くことがある。これらの化合物には共鳴周波数が 90 MHz で測定された ^{1}H NMR が多いが、現状により即した 400 MHz の情報も加えることが今後必要であろう。

5 まとめ

1982 年に構築を開始して以来、産総研の有機化合物のスペクトルデータベースはこれまでに 3 回の大きな世代交代を伴った。直接開発に中心的に携わった第1世代はプロジェクトを立ち上げ、その後のこのデータベースにとって重要な方向付けを行った。第 2 世代はウェブ公開を実現し、大型コンピュータが利用できなくなった際のデータマネージメントシステムの原型を完成した。大型コンピュータでは利用できなかった小文字への対応も開始され、化合物名称や分子式等の表記上の問題の解決が行われた。

著者らは第 3 世代にあたる。2001 年に工業技術院物質工学工業技術研究所から産総研に組織が変わった頃、ちょうど現在のスタッフを中心とした活動が始まった。前後してこのデータベースのための MS、NMR や IR の装置を更新した。スペクトル毎に別々に活動してきたスペクトル担当のスタッフが、化合物辞書を整備するスタッフと活動拠点を同じにした。このような環境を手に入れたことにより、測定スペクトルに疑義があった化合物や辞書情報の内容について確認や議論を円滑に行うことが可能となり、公開前のスペクトル情報の管理や、そこから公開用のデータを作るための内部データ管理ツールを充実させることができた。ウェブでは検索等の機能拡張を行ってきており、産業界のユーザーに留まらず、ウェブ公開によって教育への利用が多くなったことで、これまでと異なるデータ収集の方針を検討することも必要になってきた。一例がスペクトル、特に ^{1}H NMR 情報の更新である。

データベースに真剣に取り組む研究者がいなければ、データベースの活動は成り立たないのは言うまでもない。加えて、この研究者の取り組に対する組織からの支援があったことが、継続的な活動を可能とした。このデータベースはウェブで多くのユーザーに支持されており、需要があることも組織的な支援を受けられた要因である。これら研究者と組織が両輪となり、ユーザーに必要とされるデータベースを構築したことが、このデータベースが長期間にわたる

活動を継続することができた要因の一つである。限られた資源の中で、信頼性の高いデータベースの情報を発信し続けることはたやすいことではない。NMR を例にとると、化合物が分解しない条件下で速やかに測定を行うこと、測定を自動化すること、帰属作業の簡便さと正確さ向上のために、2 次元スペクトルを測定して、結合している ^{1}H と ^{13}C の骨格を確認すること等、信頼性の高い情報を効率的に収集する工夫をしているが、評価の最終確認は研究者の目が必要である。これを自動的かつ効率的に行えるような評価方法を確立することができれば、このデータベースも次の大きな変換点を迎えることとなろうと考える。

6　謝辞

研究活動を開始して以来、多くのスタッフが産総研有機化合物スペクトルデータベース（SDBS）の発展に寄与してきた。このデータベースにかかわった方々に、この場を借りて感謝の意を表したい。

参考文献

[1] O. Yamamoto, K. Someno, N. Wasada, J. Hiraishi, K. Hayamizu, K. Tanabe, T. Tamaru and M. Yanagisawa: An integrated spectral data base system including IR, MS, ^{1}H-NMR, ^{13}C-NMR, ESR and raman spectra, *Anal. Sci.*, 4, 233-239 (1988).
[2] 齋藤剛: 有機化合物のスペクトルデータベース, *産総研 Today*, 7 (1), 36-37 (2007).
[3] http://riodb.ibase.aist.go.jp/index.html(2010年7月30日現在)
[4] http://riodb01.ibase.aist.go.jp/sdbs/(2010年7月30日現在)
[5] 陳勝智: *Fragmentation and Interpretation of Mass Spectra*, 中国医薬大学, 台湾 (2010).
[6] 小川圭一郎, 榊原和久, 村田滋: *基礎から学ぶ有機化合物のスペクトル解析*, 東京化学同人, 東京 (2008).
[7] K. Hayamizu: An input tool by a personal computer for the NMR Spectral Database (SDBS-NMR), *Journal of Computer Aided Chemistry*, 2, 1-10 (2001).
[8] O. Yamamoto, K. Hayamizu and M. Yanagisawa: Construction of proton nuclear magnetic resonance database system with full spectral patterns, *Anal. Sci.*, 4, 347-352 (1988).
[9] O. Yamamoto, K. Hayamizu and M. Yanagisawa: Construction of proton nuclear magnetic resonance parameter database system, *Anal. Sci.*, 4, 455-459 (1988).
[10] 早水紀久子: インターネット上のスペクトルデータベース (SDBS), *物質研NEWS*, 37, 3 (1995).
[11] 中村徹: 化学物質リンクセンター 様々な化学物質データをワンストップで, *CICSJ Bulletin*, 25 (4), 88 (2007).
[12] 齋藤剛: 有機化合物のスペクトルデータベース SDBS, *CICSJ Bulletin*, 25 (4), 99-102 (2007).
[13] 平泉紀久子, 和佐田宣英, 田辺和俊, 田村禎夫, 柳沢勝, 小野修一郎: 化合物スペクトルデータベースシステム(SDBS)のオンラインサービス, *化技研ニュース*, 6 (1), 2 (1988).
[14] 早水紀久子, 田辺和俊, 田村禎夫, 柳沢勝, 小野修一郎: 化合物スペクトルデータベースシステム(SDBS)のCD-ROM版, *物質研NEWS*, 9, 6 (1994).

執筆者略歴

齋藤 剛（さいとう たけし）

2000 年工業技術院物質工学工業技術研究所入所、現在は産総研計測標準研究部門計量標準システム科主任研究員。有機化合物のスペクトルデータベース（SDBS）の高度化研究に従事し、現在は SDBS を統括している。NMR を利用した研究を行っており、NEDO 委託事業「ナノ計測基盤」では、NMR を利用した液中粒径計測の研究に従事、現在は、NMR を用いた定量技術の高精度化とこれを利用した SI トレーサブルな標準物質供給に関する研究に取り組んでいる。この論文では全体を統括した。

衣笠 晋一（きぬがさ しんいち）

1987 年工業技術院化学技術研究所入所、現在は産総研計測標準研究部門先端材料科高分子標準研究室長。高分子の分子特性解析をベースに、高分子標準物質、ナノ粒子標準物質の研究開発に従事している。2001 年より有機化合物のスペクトルデータベース（SDBS）の高度化研究に従事、主に赤外吸収スペクトルを担当している。この論文では全体を齋藤と共に担当した。

査読者との議論

議論1　全体的コメント

コメント（富樫 茂子：産業技術総合研究所評価部）

産総研が公開しているデータベースの中で、外部からのアクセス数が最も多い有機化合物のスペクトルデータベース（SDBS）に関して、データベース構造、データ集積、データ公開の方法論が述べられており、本誌にふさわしい研究論文と考えます。

コメント（小野 晃：産業技術総合研究所）

30 年間にわたる長期のプロジェクトに対して、その基本構想から開発、維持、公開に至るコンセプトとプロセスが分かりやすくまとめられています。この研究は第 2 種基礎研究から製品化研究にわたる広範な研究業務で、産総研のような公的研究機関にふさわしい研究成果と思います。また世界中から大量のアクセスがあることも、このプロジェクトの成功を示すものと言えます。

議論2　アクセスログの解析と公共財

コメント（富樫 茂子）

アクセスログの解析をすることでユーザーの情報がかなり得られるはずです。外国・国内の別や、大学・公的機関・企業・一般等のおよその分類ができるはずですので、加えてはいかがでしょうか。

また、文中で「無料の公開」が頻繁に強調されています。公的機関が公共財として社会に幅広く活用される情報を無料で公開することは極めて重要な役割と考えます。独立の章をたてて、この点を十分に議論していただけると、シンセシオロジーの論文として深められると思います。

回答（齋藤 剛）

アクセスログについては、アクセスするユーザーの国や、国内ユーザーの（ac や co のような）ドメインに関したアクセス状況について図を追加し説明を加えました。

「無料の公開」に関しては、私達も独立の章を設けて議論することが望ましいと考え、4.2 節の中に新しい段落を起こして無料サービスの意義に言及しました。

議論3　データベースにかかる人と経費

質問（小野 晃）

産総研の有機化合物スペクトルデータベースの構築と公開に要するコストについて伺います。データベースシステムの構築（ハードウエアとソフトウエア）、試料の入手、測定の実施、データの管理と品質の確保、ユーザー対応等にかかるマンパワーと経費は大まかにどの程度のものでしょうか。

回答（齋藤 剛）

産総研発足から2007年度までは研究者が全体のデータベースの方針等の策定、スペクトル評価等を行い、MS、IR、NMRと、化合物辞書を担当する契約職員4名の体制でデータベース構築を行い、構築したデータの公開は産総研の研究情報公開データベース（RIO-DB）のシステムエンジニア（SE）に多くの作業を依頼しています。

私達がスペクトルデータベースの作業にかかわった期間では、研究者は一年あたりの延べ作業量として、データベースシステム構築0.25名、測定0.25名、データの品質確保に0.8名、データ管理とユーザー対応を合わせて0.25名程度のマンパワーを費やし、経費は概算で一年間当たりデータベースのハードウエアシステムの構築に20万円、ソフトウエアシステム構築に150万円、試薬の入手に25万円、装置の購入費を除く測定にかかわる消耗品や装置のメンテナンス費用として180万円、データ管理に70万円程度を要しています。このほかに、データベースの公開を担当するSEに多くの作業を依頼していますが、これに関しては私達は把握できておりません。

議論4　網羅性、信頼性、緊急性のバランス

質問（小野 晃）

①データベースの構築では常に網羅性と信頼性のバランスが問題になることはこの論文でも述べられています。このデータベースの目的を、汎用の化合物を同定するための標準スペクトルデータの提供とし、データの信頼性を第一に考え、試料に関する情報と測定はすべて自己（産総研）が把握・管理できる範囲に限定するという方針を採ったと理解しました。このためデータベースの網羅性は、その達成が後回しになってもやむを得ないとし（すなわち時間が解決するという方針をとり）、開始30年後の現在では十分な量（3万種類の化合物）に達した、という理解でよろしいでしょうか。

②一方、汎用の化合物だけでなく、最近ではこの論文でも指摘されているように農薬や劇物等、社会が緊急に求めている特殊な化合物のスペクトルデータも求められているように思います。これらのスペクトルデータベースを構築して公開することは重要だと思われますが、現在世界のどこかの機関から公開されているのでしょうか。これらのデータベースはユーザーから見て十分な状況にあるのかどうかお尋ねします。

③もし十分な状況でない場合、緊急かつ大量にスペクトルデータが必要ならば、これまでの産総研の対応方針ではニーズに対して間に合わない恐れがあります。農薬や劇物等のスペクトルデータベースに関しては、その信頼性をやや落としてまでも、網羅性と緊急性を最優先にしたデータベース構築が求められるように思いますが、この点に関して著者の方々はどのような見解をもっておられるかお伺いします。

回答（齋藤 剛）

①基本的には、データベースの網羅性より信頼性を優先した結果、データ量を急激に増加することができなかった点も、活動を長期にわたって継続した結果、化合物数で3万件、スペクトル数で10万件と大規模になったことも、ご指摘のとおりです。汎用性の高い化合物は網羅できたと考えられます。

NMRに限定すると、スペクトルデータの公開にはスペクトルを測定するだけでなく、帰属も行うこととしたので、活動を行っている人的、装置的な時間制約でこれ以上データ量を増やすことが困難でした。また、ほかのスペクトルについても研究所内やほかの機関から試薬の調達を試みましたが、十分な試薬を集めるための予算が不足していた点も網羅性を達成する妨げの一因であったと考えられます。

②医薬品、毒物、農薬、汚染物質のマススペクトル・データベースがJohn Wiley & Sons社からCDと本のセットで、農薬等の環境関連IRデータベースがBio-Rad社から提供されています。劇物については、国内法にのっとった分類であり、このような形で分類されたスペクトル集はないと思います。このような分類を明示していませんが、農薬や劇物に分類される化合物のスペクトルデータは他にもあると考えていますが、ユーザーから見て十分な状況にあるとは言い切れないと考えているため、このデータベースでもこれらのスペクトルデータの整備を行っています。

③現在の体制では、スペクトルの信頼性を落とすだけでは限界があり、網羅性と緊急性に対応しきれない面もあります。これを達成するためには、緊急性の高い化合物のスペクトル情報を優先的に測定、評価して、これらのデータ公開をしていくプロジェクトを立ち上げる方法が良いと考えています。一つの選択肢として、産総研外のデータを収集する方法やスペクトル評価基準を構築して、将来的に公募形式のスペクトルデータベースへ発展させることがあります。こうすれば一定水準の品質を確保した上で、より網羅性を備えたスペクトルデータベースへ発展させることが可能ではないかと考えています。

議論5　デジタルデータと著作権

質問（小野 晃）

このデータベースの中ではデータはデジタル化して管理されていますが、ウェブに公開するときにはアナログ化し、ユーザーはデジタルデータにはアクセスできないようになっていると理解しましたが、それでよろしいでしょうか。

ユーザーがデジタルデータにアクセスできない理由には、産総研が取得したスペクトルデータには著作権があり、第三者がデジタルデータを使用したいときには著作権料を支払うことになるという理解でよろしいでしょうか。

回答（齋藤 剛）

ご質問にある、「デジタルデータ」が「スペクトルを構成するポイントが座標情報として示されたデータ」、「アナログデータ」が公開に用いている「gif画像データ」という意味で、ご指摘のとおりです。

ユーザーがデジタルデータにアクセスできないようになっている理由は、著作権や著作権料自体の観点からではなく、著作権保護の観点からです。つまり、SDBSを不当な模倣から守り、また、模倣によって第三者が不当な利益を上げるのを未然に防ぐためです。デジタルデータは加工性が高いため商業的価値が高く、もしデータが大量にコピーされればSDBSと同等、あるいはそれ以上のデータベースを簡単に作られてしまう可能性があります。これはSDBSにとっては非常に脅威です。仮に著作権侵害が認められ裁判所に訴えることができても、そのためにかなりの労力を費やさなければならないと考えます。ところで、現在公開しているアナログデータも大量にコピーされれば著作権侵害でありSDBSに脅威を与えるので、SDBS防御の立場からアクセス状況を絶えず監視しています。

一方、アナログデータであれデジタルデータであれ、第三者がデータを利用したい場合には産総研からの使用許可が必要であり、特に第三者がデータを販売したいときは著作権料を産総研に支払うことになります。これはウェブの公開における著作権の問題とは別の話になると考えます。秘守義務の関係上会社名を挙げ

られませんが、産総研発足後、大量のデータをまとめて提供したことが数件あり、いずれも契約に基づいて著作権料の支払いを受けています。また、現在でもIRスペクトルについては新規公開するごとに提供し、著作権料の支払いを受けている案件があります。ちなみに、米国の某データベースにはアナログデータ（gif画像データ）を提供した経験があります。

議論6　ほかのスペクトルデータベースとの比較

質問（小野 晃）

世界にはこのデータベースのほかにも公開されているスペクトルデータベースがあるのではないかと思います。特に有料でデジタルのスペクトルデータを企業等に頒布するサービスをしている企業はあるのではないでしょうか。それらを紹介していただき、このデータベースとの役割や特徴の違いに関してご教示願います。

回答（齋藤 剛）

まず、ウェブ公開しているスペクトルデータベースはそれほど多くありません。このデータベースのように無料で多くのスペクトルデータを閲覧可能なものは限定されます。特にこれだけ^{1}H NMRのスペクトルパターンとその化合物への帰属情報を無料で閲覧することができるスペクトルデータベースは、このデータベース以外には著者が調べた範囲ではありません。

数少ないウェブ公開のデータベースの中でまず挙げられるのは、米国国立標準技術研究所（National Institute of Standards and Technology, NIST）が提供するNIST Chemistry WebBook（http://webbook.nist.gov/）です。物理化学情報やスペクトル情報等、さまざまな情報をウェブを通して無料で利用することができます。これは産総研の研究情報公開データベース（RIO-DB）と類似しており、NISTで得られた成果を中心にそのデータを公開しているもので、公開されているスペクトルはMSが1.5万件、IRが1.6万件のほかに紫外可視吸収スペクトルやテラヘルツスペクトルがあります。ウェブを通した利用は、NIST WebBookを閲覧するために特別なソフトウエアをインストールする必要はありません。化合物を検索した結果から、スペクトルやそのほかの情報を参照する形で構成されています。化合物リストを把握しておらず推測ですが、一般的な試薬は多く収録されており、本スペクトルデータベースと同様公的機関としての役割を担っていると考えます。この一方で、MSデータはNISTで評価したスペクトルを中心に、米国国立衛生研究所（NIH）と米国環境保護庁（EPA）のデータを合わせてNIST 08 Mass Spectral Library（2008年に発表されたもので、旧バージョンは2005年に発表された）としてパソコン単体で利用するよう販売されています。データはウェブで公開されているデータ数よりはるかに多く、約19万の化合物に対して22万件のスペクトルが収録されています。旧工業技術院の時代に共同研究を行った際に、このデータベースのMSデータも数多く登録されており、このデータもこのライブラリのデータの一部になっていると思われます。このデータベースは有償で配布されており、多くの質量分析装置に搭載して、質量ピークのパターン検索に利用されています。

SpecInfoは^{1}H NMRが9万件以上、^{13}C NMRが30万件以上、このほかの核種のNMRやIRやMSも収録されているスペクトルデータベースで、ウェブをとおして有料公開を行っていますが、NMRスペクトルが2006年に更新されたのを最後に、データの更新がされていないようです。

国内では、分散型データベースで基礎代謝、植物二次代謝標準物質の高精度精密質量スペクトルを対象として、ウェブでデータ公開を行っているMassBank（http://www.massbank.jp/）があります。2010年12月1日の時点で、19研究機関から約3万スペクトルがウェブで無料公開されています。登録されるデータが、このデータベース（SBDS）とは異なり植物二次代謝物質と専門性の高い領域をターゲットに高分解能MSスペクトルを集積しており、検索等に利用するツールをクライアントパソコンにインストールする事で、データベースの検索や、スペクトルの拡大等を含めたデータ参照、そしてデータ登録を行うことができるようになっています。2006年度から、（独）科学技術振興機構バイオインフォマティクス推進事業の研究課題「メタボローム・マススペクトル統合データベースの構築」で構築を行っているもので、このデータベースも開発当初はこのようなプロジェクトで基礎を固めたことにかんがみると、このプロジェクトが修了した後どの様な形にMassBankが発展して行くか期待しています。

ウェブ公開ではない形で提供しているものにBio-Rad社のスペクトルデータベースがあります。かつてのSadtler社のデータを中心に、SpecInfo、NISTのMSやこのデータベースのNMR等のデータを収録しており、検索したり物質の同定に活用したりする独自の検索ソフトウエアである「KnowItAll」と合わせてパソコン単体にデータをインストールして利用する形で販売されています。登録されているスペクトル概数は、^{1}H NMRが5万件、^{13}Cが43万件、MSが19万件、Ramanが7千件、IRが23万件です。^{1}Hと^{13}C NMRにはこのデータベースのスペクトルのデータがそれぞれ約1.3万件と1.1万件登録されています。それぞれのスペクトルのパターン検索を利用することが可能で混合物のスペクトルマッチング等、複雑な検索も行うことが可能なソフトウエアを搭載しています。

一方、試薬会社のSigma-Aldrich社が、「Aldrichスペクトルライブラリー」を販売しています。これにはNMR、IRとRaman合わせて延べ5万物質以上が登録されており、パソコン単体で利用する形態と、スペクトル集の書籍としての形態で販売されています。

ウェブ公開するデータベースとパソコン単体で駆動されるデータベースとを比較すると、それぞれ一長一短があります。例えば、ウェブで公開するデータベースの利点としては即時性が挙げられます。このデータベースはデータの追加、更新作業を比較的簡易に行うことが可能であり、年に2回のデータ更新を行っており、常に最新のデータをユーザーへ供給しています。ただ、上に示した多くのデータベースは、スペクトルの追加、更新作業をこのような頻度で行っていないようです。一方、パソコン上で単独に駆動されるデータベースは利便性に勝り、例えばスペクトルのパターン一致検索ができる等、ユーザーへの利便性が高いものとなっているようです。