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 Shinichi 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 industry and society are objects of compilation. On the other hand, the computer system used for database management and the ways for data release has 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 Introduction

The reliable analysis of chemical substances is required not only in industry but also in various situations in society. The nuclear magnetic resonance (NMR), the infrared (IR) and the mass spectroscopy (MS) give powerful and essential information in the identification of an organic compound. The measurements and the analyses of spectra need to be performed whenever a new chemical is developed, unknown materials are analyzed or chemicals are identified.

In general, identification of a material from spectra by comparing and matching with those obtained as a standard spectral data is one of the most reliable methods. Since this approach has been used in a variety of applications, the role of such standard data and their database is important. Spectral database system for organic compounds, i.e. SDBS, of the National Institute of Advanced Industrial Science and Technology (AIST) was started in 1982 as a project of the former Agency of Industrial Science and Technology. Since the start of this database project, two basic concepts of 1) collection of standard spectral data that are reliable and 2) compilation of more than two kinds of spectra for each compound have been maintained. In other words, its basic concepts are acquisition of as many as six spectral data of MS, ^{13}C NMR, ^{1}H NMR, IR, Raman and electron spin resonance (ESR) spectra for a compound by ourselves (AIST), and management and maintenance of the associated chemical information for the compound[1].

During almost 30 years of the database activity, collection of

Raman and ESR spectra have been discontinued. Currently, activities on four kinds of spectral data of MS, IR, ^{1}H NMR, and ^{13}C NMR have been continued along with the management of compound information, and these spectral data and chemical information are disclosed to the public[2].

In 1997, we started to release the database contents to the public through the Web[3] by a project of the former Agency of Industrial Science and Technology[4]. The total number of compounds and spectra disclosed as of April, 2010 is about 33,000 and about 100,000, respectively. Figure 1 shows the number and fraction of each spectrum. The main users of the present spectral database are those who access through the Web. Since the database has been disclosed to the public through the Internet, it has had many accesses. The average number of page views per day during the last three years exceeds 100,000 times, which is exceptionally high
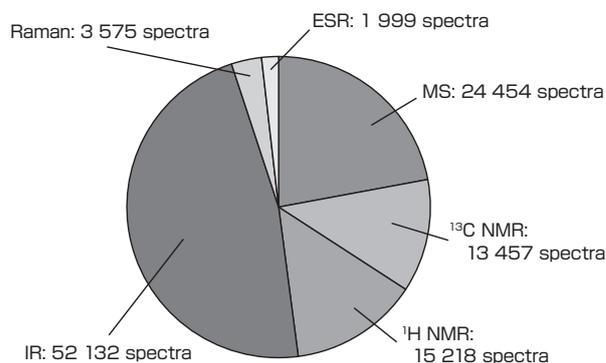

Fig. 1 Fractions and the number of spectra open to the public by the spectral database for organic compounds (SDBS) of AIST as of April, 2010

National Metrology Institute of Japan, AIST   Tsukuba Central 3, 1-1-1 Umezono, Tsukuba 305-8563, Japan   ∗ E-mail : takeshi.saito@aist.go.jp

among databases open through the Web by the "Research Information DataBase (RIO-DB)" that is operated by AIST. The total number of page views since the database was disclosed to the public exceeded 300 million at the end of fiscal year 2009. The transitions of the number of accesses to the database and the number of spectra compiled to the database per fiscal year are indicated in Figs. 2 and 3, respectively. The number of accesses has been increasing every year in the last ten years because of a remarkable expansion of the usage of the Internet and the recognition of this database among Internet users. Requests for using our spectral data for textbooks[5], reference books[6], and as materials for examination problems, etc. come often from users; there are users who point out mistakes found in our database as well.

Figure 4 shows the scenario of the development of this database. This figure indicates a list of various elements which compose this database, relationships between each of the elements and the main characteristics of the database such as the basic structure, completeness, reliability and usability. Important elements for the operation of this database are
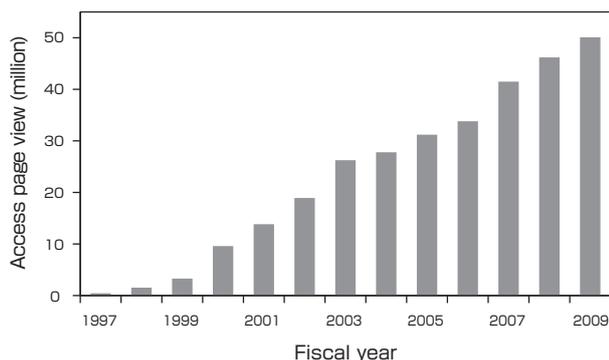
shown here. A method of how respective elements have been integrated to this spectral database is described in the following chapters.

2 Structure of database

2.1 Importance of basic structure of database
This database has taken a structure that allows two or more kinds of spectra to be accessed for each compound. In order to achieve such a structure, this database was designed and built as seven independent databases: a compound database, which has been called a compound dictionary, and six different spectral databases. All of them were integrated mainly under the compound dictionary as shown in Fig. 5.

To do this work smoothly, several kinds of management numbers were prepared. These are a bottle number which is given to a unique bottle of a chemical reagent, a spectral management number which is assigned to each individual spectrum, a spectral code which is only given to the spectrum which is compiled in the database, and an SDBS compound number (called SDBS number within our databasing group)

Fig. 2 Yearly transition of the number of access page views since opening to the Web of the spectral database for organic compounds (SDBS) of AIST
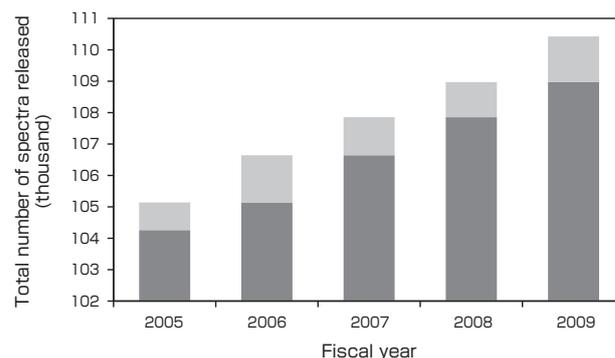

Fig. 3 Transition of the number of spectrum open to the public for the last five years for the spectral database for organic compounds (SDBS) of AIST
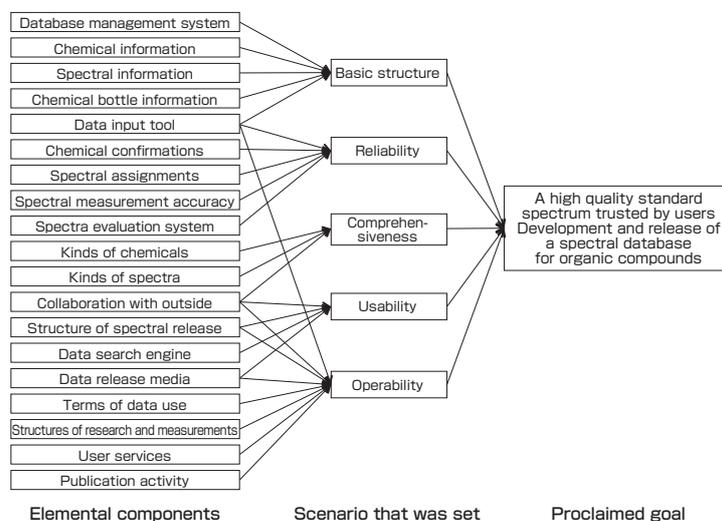New spectral data released each fiscal year are shown in faint color.

Fig. 4 A scenario of a construction of database and data release to the public for the spectral database for organic compounds (SDBS)

which is a unique number for each chemical identity. The spectral code for each kind of MS, IR, ^{13}C NMR and ^{1}H NMR spectrum has been managed individually. Thus the database for a single kind of spectrum and chemical information can be separated as an independent database. Compound and spectral information corresponding to these numbers are managed as relational databases independently, and these numbers made our work proceed smoothly.

In particular, adopting the SDBS compound number that is a unique identification to a compound is a characteristic of this database. This number is not just a management number of the compounds. This number enables the compound dictionary to be independent from a spectrum database. As a result, the flexibility of changing the compound dictionary was maintained. Thus operation of useful compound information has been possible up to now. The founders of this database had had rich experiences in constructing databases while in the gas chromatographic data committee, the infrared data committee, and the NMR data subcommittee during the era of the former Agency of Industrial Science and Technology. Based on such experiences, the SDBS compound number was adopted to this database and made the database functional for some 30 years after it was established.

In principle, all compounds have been obtained and their spectral data have been acquired by ourselves. The unique bottle number was assigned to each chemical bottle obtained. An SDBS compound number was assigned to each compound. Each sample identified in a bottle had been carefully checked for avoiding assignment of two or more different SDBS compound numbers to an identical chemical compound. This number was assigned only when the sample was evaluated as a new chemical compound in
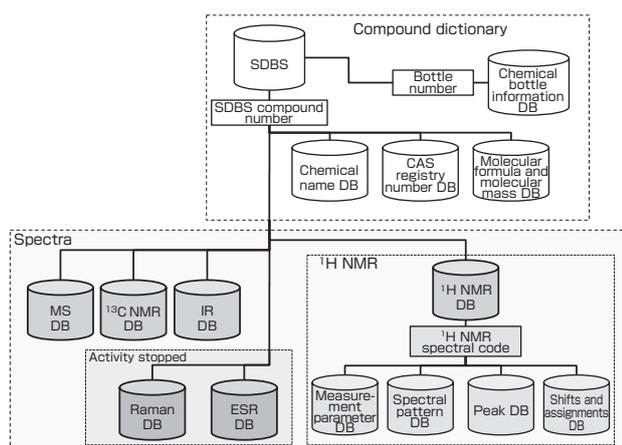

Fig. 5 Structure of the spectral database for organic compounds (SDBS)

SDBS in this figure contains information of the SDBS compound number that is the unique number for each compound, the number of elements for the compound, and other information. The structure of ^{1}H NMR database is shown on behalf of each spectral database. All information is correlated to the SDBS compound number via a relational database.

the database. This work might not have been a tough task in the early days of this database, however, as time passed, the task became more difficult. In the past few years, we encountered questions and problems in the checking process. The assignment of a SDBS compound number to a unique chemical became a more complex task and required longer time compared to the early days. This is because compound structures and thus their names have become more complex, and the number of compiled compounds has become large extending to some 30,000 different chemical compounds. For solving this problem, first we collected as much information of a chemical compound as possible, then individual information was searched through our compound database. Deeper consideration by several chemists was made only for those compounds with possibilities of having already been compiled. This enabled us to spend more time to acquire and evaluate the spectral data. As a result, the problem of assigning different SDBS compound number to two or more identical compounds has occurred less frequently and full supervision by chemists has become unnecessary. In recent years, SDBS compound assignment process has become smooth.

2.2 Decision of the database operation and renewal of the database platform

The operation of this database was started with a mainframe computer in the1980's. The choice of such a computer made sense because the first Japanese Windows computer, NEC PC-9800, was just released about the same time. However, operation of this mainframe computer (FACOM MSP) ended in March 1999 due to the decision by the former Agency of Industrial Science and Technology. It was needed to decide whether activity of the database would be continued with another mainframe computer or a personal computer (PC), or terminated altogether. At that point, we decided to continue our activities and to adopt Windows PC, and a new data input tool was developed on the PC[7]. With much of the computer software, it was difficult to shift platforms from MS-DOS to Windows computers, and the majority of the system had to be rebuilt. This database, however, was successfully and completely transferred to a different platform of its operation. If we had remained with the mainframe computer at that time, our system might have been in trouble in adopting many tools for assisting data input to the database system. The data collection and handling became greatly easy by adopting Windows PC as a platform for the data management.

3 Selection of data to be collected

3.1 Strategy for selecting compounds

This database was designed as a useful tool for compound identification. Data collections were mainly focused on spectral data of commercial reagents that were frequently used by many people. Figure 1 shows the number of spectral

data released through the Web. There are many more spectra that had not been released to the public. The total number of bottles of chemical reagents exceeds 39,000.

Among these, more than 10,000 reagents have been offered free of charge by Tokyo Chemical Industry Co. Ltd., from which the most abundant chemicals have been supplied. Therefore, although the selection of chemical reagents partially has followed this company's policy in development of reagents, it has indirectly reflected our user's needs. In the development of a new material by chemical synthesis or other methods at research and development departments, a starting material is in many cases a commercial chemical reagent. Thus, the support we have been given by the company has been valuable to us.

Since 2001, our strategy for a spectral collection has focused on pesticides and deleterious substances. Collection of much spectral information on regulated chemicals is an important function of public research institutes like AIST. Thus the number of spectra collected for such substances has been slowly increasing. Recently, the concern for the safety of food has been increasing, which enhances the need for such information. It is important that our strategy focus on collecting the spectral data of pesticides and regulatory chemicals.

3.2 Selection of visual data form (digital data)
The most important decision for the selection of data format was made at the early stage of this database. Although it is not surprising now, this database chose to collect all spectral information in digital coordination format on a computer since the activity was started. In the 1970's, spectral information was collected more often in a data book format. Although it was recognized that the handling of spectral data would become easier by making them digital, limitation of the computer memory prevented this. Due to the limitation of the computer, digitization of the data often resulted in the loss of information. As a result, the analog data format recorded on paper was still the majority[1] at that time. For example, NMR data was composed of several tens of thousand points of data. Digitalizing such data about thirty years ago must have been a big decision because of the limitations in disk and memory capacity. An achievement of such a system would have been extremely difficult if there had been no mainframe computer operated at the former Agency of Industrial Science and Technology at that time. Under this condition, management of the spectral database required not only to concentrate on accumulating spectral data, but also to find a creative way to minimize the data points. This system was the world's first ^{1}H NMR spectral database with digital coordinate data of a collected spectrum[8]. We adopted compression of data size by collecting data which represented only peak areas. For the ^{13}C NMR, values of peak positions, their intensities and their peak width at half high were recorded. From these data, all

spectra were reconstructed with the assumption of all peaks as the Lorenz function. For the IR and Raman, coordinate data of the spectral points were collected. For the MS, the mass numbers and the signal intensities were collected. For ESR, each point of spectral data was digitalized. Some of the data was reconstructed from paper data by using a curve reader. ^{1}H NMR has a capability of spectrum simulation by using chemical shifts and spin-spin couplings[9]. After AIST was established, all digital data including the peaks and noise have been collected for ^{13}C NMR and ^{1}H NMR. Users can recognize the strength of the peak signal compared to the noise level. In 1997, this database became disclosed to the public through the Web by the former Agency of Industrial Science and Technology. If all the data were not collected digitally, there would have been problems for converting old analog data, and many data might have had to be recollected.

3.3 Balance between quality and quantity of data; stick to the high quality data
The spectral database consists of data acquired, evaluated and compiled in our institute with some exceptions in ESR and ^{1}H NMR spectra. This is the most reliable way to keep the quality of the spectral data. This makes quality of the data reliable. However, on the contrary, the number of accumulated spectral data has become limited. To cover a wide variety of data is one of the important elements of a database. How we balance the two different concepts for the data collection, i.e. quantity and quality of spectral data makes for serious argument. Our first decision was to take a strategy of collecting reliable standard data. On this basis, the quantity of the data would be increased as a result of data accumulation over a long period of time.

The criteria for keeping quality and reliability of spectral data and for accumulation of data were established. For example, tetramethylsilane (TMS) was not only used as a chemical shift standard for NMR spectra but also its line width was used for a criterion of spectrum resolution. When the peak resolution of the TMS peak was sharper than the criterion, the resolution of the spectrum was determined well even if the resolution of the peak from a compound showed poor resolution. This was considered as the nature of the compound giving such a spectrum, and not caused by the bad experimental condition. For IR, no interference noise, no water peaks, or no surge in baseline were the criteria. A criterion for each spectral data evaluation was established by respective researchers in charge of the spectrum.

3.4 Policy of data registration
This database only compiles unique spectral data. In other words, when several spectra for a compound have been acquired with identical conditions, only the spectrum of the best quality is compiled and released to the public. For MS, a direct sample injection method was adopted for the measurement. Therefore, each compound had a unique

measurement condition. Only the best quality MS spectrum was compiled for a molecule. For IR, different conditions of measurement such as KBr plate method and Nujor method for a solid sample was adopted. In such cases, spectra with the best quality and reliability in each condition were compiled. For ^{13}C NMR, ^{1}H nucleus was decoupled from the spectrum which made the spectrum simple singlet peak for each carbon for most of the cases. Only one ^{13}C NMR spectrum was compiled for a compound.

For ^{1}H NMR, a resonance peak pattern depends on the resonance frequency. This makes spectral pattern for a given sample solution to be also dependent on resonance frequency. Many ^{1}H NMR were acquired at resonance frequency of 90 MHz at the beginning of this database activity. When a molecule had a complex structure and it was hard to make chemical shift assignment at this frequency, a higher frequency of 400 MHz, which simplified spectral patterns, was also adopted for such a molecule. Simulation of spectral pattern was also important. This database had a capability of simulating spectrum observed at different resonance frequencies. For fulfilling this capability, chemical shifts and spin-spin coupling constants for a molecule were also compiled in the database independent of spectral patterns.

For ^{1}H and ^{13}C NMR, chemical shift assignments were compiled in addition to the information above. Especially for ^{1}H NMR, since the resonance peak pattern depends on resonance frequency, it was indispensable to provide a universal property of chemical shift assignments. If there was no such data compiled in this database, two ^{1}H NMR spectrum obtained at different resonance frequencies cannot be compared. The chemical shifts and their assignments were the most valuable information for the ^{1}H NMR spectral database.

Information on a chemical compound was compiled as much as possible. It is usual that the more complex the structure of a compound is, the more chemical names or abbreviations are assigned to it. We thought that it was beneficial for users of this database if they could easily find a compound with various names, and therefore, much chemical information was compiled encyclopaedically.

3.5 Kinds of spectra collected
In the early days, six kinds of spectra were collected. Currently, four out of the six kinds are still actively collected (Fig. 5). The spectra used in analysis in the 1980's were not limited to those six kinds of spectra. Spectra such as ultra violet-visible spectral data were not incorporated in this database. The selection of the six spectra for this database is thought to have depended on instrumentations and researchers in our institute at that time. The reasons for the decision to terminate the activities of Raman and ESR spectra were not only based on the instrumentation

and the researcher issues but also on the then supposed relatively low potential demands from users of the database. However, the demand for Raman spectral data has now expanded academically and in industry. Based on this point, this database may not cover enough to respond to the user's needs. On the other hand, MS, ^{13}C NMR, ^{1}H NMR, and IR have had huge demands from the 1980's up to the present days. After this database was released to the public, the demands of users can be evaluated by the number of accesses to spectral data (Fig. 2).

4 Policy of data release to the public

4.1 Releasing data on the Web
In 1997, spectral data of MS, ^{13}C NMR, and ^{1}H NMR was released from the website of AIST, and IR and ESR followed a year later[10]. Currently, six kinds of spectral data, including Raman spectra, are opened to the public through the Web. When our data were opened to the public for the first time, people had started to use web browsers such as NCSA Mosaic and Netscape. Although more and more people started to use the Web, the internet line was slow and functions of the browsers were insufficient compared to the current situation. Therefore, it was important to display the information efficiently on the Web. As a result, we decided to use GIF image format for displaying spectral patterns and structures of molecules which placed the least load on the internet line. The speed of the web access has increased rapidly in Japan, and therefore, it may not be applicable to other particular parts of the world. Therefore, this format is still kept in our system.

Another reason to use the GIF format is to protect the data. In other words, image display format can prevent illegal loss of coordinate data or intellectual property of our database. It is easy to reproduce a high resolution spectrum using digital data. On the contrary, it is not possible to create spectral data with higher resolution than the original image when it is reconstructed from the image. We have had several experiences of systematic data download activities for taking all or at least the majority of our data. The protection plan that we took worked for preventing the coordinate data to be inappropriately taken. In the future, with sufficient protection against illegal accesses, spectral expansion functions based on the coordination data can be provided on the Web.

We chose English for information on the Web. This choice was possible because the collected compound names were in English. The other information, or the spectral data, did not depend on any languages. At present, when users access with the Japanese setting computers, the database gives frame information in Japanese.

To increase the users' convenience, links to other databases disclosed on the Web have been created for the information

that cannot be maintained by ourselves. Since 2006, the links to online catalog of Tokyo Chemical Industry Co., Ltd.,[11] and Chemical Materials Link Center operated by Japan Science and Technology Agency[12] have been maintained. Components, such as a search system for Japanese compound names and chemical structures which have not been covered by us, have been supplemented by the links.

One of the advantages of data release by the Web is its easiness of data management in a lump. It became possible to provide all users with equal services at a time by updating data in the database server. A feature different from other researches is the fact that the comments from the users come directly to us through the Web.

Before the data was opened to the public through the Web, an online access was possible since 1989[13] and the databases were sold in a CD-ROM medium with both data and their search engine from 1991[14]. It was only specific users of some tens of domestic users that were able to use this CD-ROM medium. In this format of distribution, contents could be stored for a long time. However, it was limited to the data collected up to the point the CD-ROM was created. Updating data contents and renewal of data managing software were difficult for all the users. Moreover, the service was offered to limited users who owned the CD-ROM. However, having enabled search and display of data by CD-ROM that operated on MS-DOS, it provided an opportunity of exercise for the present format of data release through the Web. This was similar to the pioneering works before the development of this database that helped to create the appropriate design of the database. Opening to the public with CD-ROM might have been an important project for opening this database to the public through the Web.

4.2 The analysis of users and the role as public resources

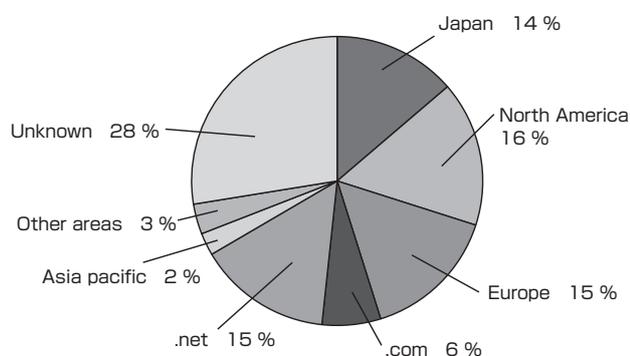For the analysis of the database users, an access log in the

fiscal year 2009 was analyzed and classified by country identification codes. Figure 6 indicates the data. Of more than 50 million page views, accesses from domestic users were about 14 %. The most accesses came from the North America region. The domains that cannot be assigned to specific regions such as ".net" and ".com" were classified independently. Focusing on the accesses from Japan, shown in Fig. 7, the most accesses came from the users of ".ac.jp" or those accesses from academic institutes, and the accesses from ".ne.jp", or network providers and ".co.jp" or general enterprises follow the academic users. Users accessed through the ".ac.jp" and ".ne.jp" domains have intense seasonal variations in the number of accesses. The accesses of March and August were less than half of June when there were the most accesses. On the other hand, roughly the same level of accesses came from the ".co.jp" domain throughout the year. The seasonal dependency of access was thought to be caused by summer break and the end of the scholastic year of students. The access tendency of users through the network providers resembles that of academic institutes. Therefore, the majority of users accessed through the ".ne. jp" were suggested to be also students; the database has been used by many students.

This database is used by various users, and it is one of the public resources that a public research organization such as AIST provides. In general, database compiles much information, and demonstrates its strength by retrieving necessary information efficiently. Many resources and much time are needed for the development and maintenance of such a database, and this database is not an exception. The database users may become limited if we ask the users for the equivalent amount of the development and maintenance costs. This database takes a role of giving various people who are in industry as well as those just starting to learn the usage of spectral data an opportunity by showing it as free accessible

Fig. 6 Fractions of regional domains of users who accessed this database in the fiscal year 2009
Domains ".com" and ".net" are counted independently of the unplaceable domains. The unknown indicates users whose access domains cannot be specified, for instance, those who accessed to the database with only IP addresses.
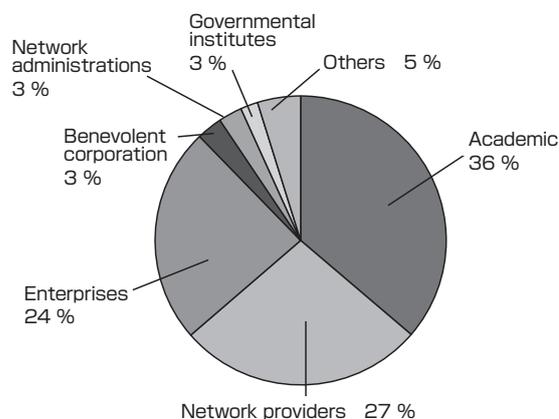

Fig. 7 Fractions of domestic domain of users who accessed this database in the fiscal year 2009
Domains of academics, network providers etc., enterprises, benevolent corporations etc., network administrations, and governmental institutes correspond to ".ac.jp", ".ne.jp", ".co.jp", ".or.jp", ".ad.jp", and ".go.jp", respectively. Other ".jp" domains are added up as "others".

public goods. Many companies can also use our spectral information in their chemical analysis without constructing their own database. The cost of the chemical analysis is, in this way, reduced. This database plays the role as an intellectual foundation that supports industry. This database has also been used by both domestic and international users for educational purposes widely for understanding the use of the spectrum. 36 % of accesses from the domestic users comes from academic sites, and there are many requests for using spectral data in text books and materials for training. From these facts, this database contributes greatly to the whole society.

4.3 Correspondence to comments from the users

We have received many comments from users by e-mail. While most of the research activities are evaluated by their research publications, this database is directly evaluated by the Web users from all over the world. Comments from the users are one of the outcomes of their database evaluation. We think it is important to catch the comments with sincerity and to use them in order to ascertain the direction and further development of the database.

The comments are classified into applications for permission to use the spectra in other materials, and technical questions. Many e-mails of appreciation of our work are delivered, which encourages us.

Examples of the technical indications are incorrect assignments of an NMR spectrum and problems related to its acquisition condition. When a comment indicating our mistakes in spectral data is received, we re-evaluate the data closely at once. When we cannot make a judgment at this point, we may acquire the spectrum of the compound again. The data will be corrected when we come to the conclusion that the user's comment is correct after our evaluations. When we reach a conclusion that our data is correct, the data will be kept disclosed. In the mean time, we explain our decision and reasons for the conclusion to the user. When needed, we obtain the compound for re-evaluation. However, when this is not possible, the data may be withdrawn.

We accept applications for the permission to use the GIF spectral data released in the Web in other materials as much as possible. We think each comment is an indication of the high quality of the data that has been disclosed in this database. We think it is important to maintain a system that can correspond to such a comment promptly.

From the analysis of the data access log, we have already indicated that we receive many accesses from educational sites. Inquiry concerning a spectrum, especially for ^{1}H NMR, of a compound that frequently appears in a textbook is sometimes received. ^{1}H NMR spectrum of such a compound is often acquired at 90 MHz. It is necessary to replace such a spectrum to one acquired at 400 MHz which is suited to the current state.

5 Summaries

Since the beginning of its construction in 1982, this spectral database of organic compound of AIST has so far undergone three generation changes of researchers. Researchers involved in the first generation had started up and made the direction of this database. Those in the second generation released the spectral data to the public through the Web, and completed a prototype of the data management system on a personal computer. Correspondence to the small letters that could not be used on the mainframe computer were started. Problems in notations of compound names, molecular formula, and others were solved.

We are the third generation researchers. Our activities started with the reorganization of our research organization from the National Institute of Materials and Chemical Research of the Agency of Industrial Science and Technology to AIST in 2001. Instrumentations for MS, NMR and IR were renewed. The staff in charge of each spectrum who used to work separately at different places worked together at one site with a staff member who had been maintaining the compound dictionary. With this environment, confirmation and discussion of spectral data and information of dictionary contents could easily be performed when a doubt in a spectral data occurred. An internal data management system, which has a capability of creating data for disclosure, was developed to maintain spectral data and chemical information. Functionalities such as the search engine have been expanded in the Web page. By releasing the data through the Web, academic users have increased, and the users are not limited to industry which had been our primary users. We need to consider development of a new policy of data collection. One of the examples is a spectrum especially the update of ^{1}H NMR information.

It is needless to say that the activity of the database does not work without a researcher who seriously works on it. In addition, continuous activity has been possible because of support from the organization to the researcher. This database has been supported by many users through the Web. Such a demand enabled us to receive support from inside the institute. The fact that the researcher and the organization have become the two wheels is one of the reasons for long term activities of this database. It is not easy to keep releasing highly reliable information to the database with limited resources. In NMR spectral activities, we have collected reliable information effectively by acquiring spectral data soon after its sample preparation to avoid the sample degradations, adopting automated spectral acquisition system, and acquiring two dimensional spectra and obtaining ^{1}H and ^{13}C skeleton of a molecule to increase accuracy of

spectral assignments. Final data evaluation is performed by the researcher. When an automatic and more efficient evaluation method could be established, it would become the next big conversion point for this database.

6 Acknowledgements

Many staff members have contributed to the development of Spectral Database for Organic Compounds (SDBS) since its activity started. We want to express our gratitude to the people who have contributed to this database.

References

[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] T. Saito: Yuki kagobutsu no supekutoru detabesu (Spectral database for organic compounds), *Sansoken Today*, 7 (1), 36-37 (2007) (in Japanese).

[3] http://riodb01.ibase.aist.go.jp/sdbs/ (As of April 1, 2011)

[4] http://riodb.ibase.aist.go.jp/index.html (As of April 1, 2011)

[5] S. C. Chen: *Fragmentation and Interpretation of Mass Spectra*, China Medical University, Taiwan (2010).

[6] K. Ogawa, K. Sakakibara and S. Murata: *Kisokara Manabu Yuki Kagobutsu No Supekutoru Kaiseki*, Tokyo Kagaku Dojin, Tokyo (2008) (in Japanese).

[7] K. Hayamizu: An input tool by a personal computer for the NMR Spectral Database (SDBS-NMR), *J. Comp. Aid. Chem.*, 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] K. Hayamizu: Intanetto jo no supekutoru detabesu (SDBS) (Spectrum database (SDBS) on the Internet), *Busshitsuken NEWS*, 37, 3 (1995) (in Japanese).

[11] T. Nakamura: Kagaku busshitsu rinku center, samazamana kagaku busshitsu deta o wan sutoppude (Chemical substance link center, various chemical substances data by one stop), *CICSJ Bulletin*, 25 (4), 88 (2007) (in Japanese).

[12] T. Saito: Yuki kagobutsu no supekutoru detabesu *SDBS* (Spectral database for organic compounds), *CICSJ Bulletin*, 25 (4), 99-102 (2007) (in Japanese).

[13] K. Hiraizumi, N. Wasada, K. Tanabe, Y. Tamura, M. Yanagisawa and S. Ono: Kagobutsu supekutoru detabesu shisutemu (SDBS) no onrain sabisu (On-line service of compound spectrum database system (SDBS)), *Kagiken News*, 6 (1), 2 (1988) (in Japanese).

[14] K. Hayamizu, K. Tanabe, Y. Tamura, M. Yanagisawa and S. Ono: Kagobutsu supekutoru detabesu shisutemu (SDBS) no CD-ROM ban (CD-ROM version of compound spectrum database system (SDBS)), *Busshituken News*, 9, 6 (1994) (in Japanese).

Authors

Takeshi SAITO
Joined the National Institute of Materials and Chemical Research of the Agency of Industrial Science and Technology in 2000. Engaged in research for development and advancement of the Spectral Database for Organic Compounds (SDBS). Currently, member of the National Metrology Institute of Japan, AIST, where he has made extensive use of NMR in his research; he is a leader of SDBS project. Worked with a project conducted by the New Energy and Industrial Technology Development Organization (NEDO) on platform for measurement at nano-scale, focusing on measurement of particle diameter in liquid using NMR. Currently, working for improvement of precision and accuracy of quantitative analysis using NMR. Also works on the development of SI traceable reference materials using the NMR technique. Saito put together the overall of this paper.

Shinichi KINUGASA
Joined the National Chemical Laboratory for Industry, the Agency of Industrial Science and Technology in 1987. Section chief of the Polymer Standards Section at National Metrology Institute of Japan, AIST. Engaged in research on polymer characterization and development of polymer and nano-particle reference materials. Since 2001, in charge of IR spectrum of SDBS. Kinugasa also put together the overall of this paper with Saito.

Discussions with Reviewers

1 General evaluation
Comment (Shigeko Togashi, Evaluation Department, AIST)

The methodology of the database structure, the data accumulation, and the data release is described for the spectral database of organic compounds (SDBS) whose number of accesses from the outside is the most among the databases opened to the public by AIST, and the paper is considered a suitable research paper for this magazine.

Comment (Akira Ono, AIST)

The concept and processes of the database have been plainly brought together from the basic plan to the development, maintenance, and data release for a long-term project of 30 years. I think this research has brought about a suitable result for a public research organization like AIST in a wide research area covered from the *Type 2 Basic Research* to the *Product Realization Research*. Moreover, it can be said that the success in this project is shown by the enormous number of accesses from all over the world.

2 Analysis of access log and function as public goods
Comment (Shigeko Togashi)

Considerable user information should be obtained by analyzing access log of the database. The classifications of international or domestic users, of academic, public organizations, enterprises, or general users, etc. must have been recorded in the log. I think adding such classifications would be beneficial to the reader.

It is frequently emphasized in the paper that the database is "open to the public free of charge". I think releasing free useful information to the public that can be used widely as public goods is a crucial role for the public research institute. I think, as a paper of *Synthesiology*, it is beneficial if an independent chapter discussing these points is created.

Answer (Takeshi Saito)

About the access log, a figure which summarizes countries of accessing users and domains such as ".ac" and ".co" of domestic users is added.

We agree it is beneficial to discuss the topic of "open to the public free of charge". We created a subchapter, 4.2, and discussed the meaning of free services.

3 Persons and expenses for the database

Question (Akira Ono)

I would like to ask about the cost of development and data release of the spectral database for organic compounds of AIST. Would you estimate roughly costs and human resources spent in the development of hardware and software, sample purchases, acquisition of spectra, data maintenance and their quality control, and user support for the database, respectively?

Answer (Takeshi Saito)

Between years 2001 and 2007, the strategy of this database operations and development, and evaluation of the spectral data were charged to two researchers. Four contract staff were employed for acquisition of MS, IR and NMR spectrum data and maintaining chemical dictionary data. Each person was also responsible for maintaining the disclosing data. All the disclosure processes which opened the data through the Web, were maintained by the system engineers (SE) of the Research Information Data Base (RIO-DB) of AIST. Estimation of the total work load per year as a researcher was 0.25 person for the database system construction, 0.25 person for the spectral measurements, 0.8 person for the quality assurance, 0.25 person for the data maintenance and user support. Looking at the budget, 200,000 yen for constructing the database hardware, 1.5 million yen for the software construction, 250,000 yen for obtaining chemical compounds, 1.8 million yen for the consumable items and the maintenance of instrumentations, and 700,000 yen for the data maintenances were roughly spent each year. Besides this, we asked the SE to do much work, but we cannot estimate the cost of work done by them.

4 Balance of comprehensiveness, reliability and urgency

Question (Akira Ono)

(1) It has been described in this paper that it is important to make a balance between comprehensiveness and reliability of data for database construction. I understood that the primary objective of this database was to focus on compiling and offering standard spectral data to help identify compounds that were widely used. I also understood that you took a policy to limit information and measurements on compounds to the range which your group (AIST) could grasp and control. My understanding of this point was that you took a policy of taking the reliability of the data over the comprehensiveness so if the achievement of the comprehensiveness was postponed, it was considered unavoidable (or took a policy of "time would solve the problem of comprehensiveness"). Thirty years after the start of this activity, the database has reached a sufficient number of spectra (30,000 compounds). Is this a correct understanding?

(2) I think the spectral data for special compounds such as pesticides and deleterious substances are requested urgently by our society. It seems to me to be important to construct a spectral database for these and release the data to the public. Is there any organization in the world which releases such data? I would like to ask whether the current situation of such a database is satisfactory to the users.

(3) If it is not satisfactory, the current policy of AIST may not be speedy enough to cover a large number of spectra in a short period of time. I think spectral information of pesticides and deleterious substances need to be covered more comprehensively and rapidly even if you lose some reliability. I would like to ask how the authors think about this point.

Answer (Takeshi Saito)

(1) It is true that as a result of having given the priority to reliability over comprehensiveness, it was not possible to increase the quantity of data rapidly. As a result of having actively compiled the data for a long period of time, the database now contains more than 100,000 spectra from more than 30,000 compiled compounds. We think the compounds that are widely used have been covered by now.

When limiting it to NMR, increasing the data bulk and speeding up the data release had become difficult and the work load had almost reached the limit of capacity of our human resources and instrumentations. We not only acquired the spectra but also assigned them for data release. We think another reason obstructing comprehensiveness of spectral data other than NMR was budget that was too limited to collect enough compounds for the data acquisition.

(2) The mass spectral database of medicine, poison, pesticides, and contaminants is offered from the John Wiley & Sons Co. as a set of CD-ROM and a paper book format, and the IR spectral database of pesticides and environmental materials is offered from the Bio-Rad Co. I do not think a database of spectral data of compounds classified as deleterious substances exists because it is a classification based on Japanese law. We believe, although there is no such classifications, many databases cover such compounds in their data entry. However, as we think the situation is not satisfactory to the users, this database will keep collecting such spectral data.

(3) There is a limit in our current resources for compiling urgent data quicker just by cutting down the reliability of the spectrum. For achieving this, we think a project with priority on acquiring, evaluating and releasing the spectra of such compounds that have high urgency is useful. Another way to achieve this is to collect spectral data from people all over the world as an open data recruit system. To make this possible, we have to build up at least a standard spectral data format, data evaluation criterion, and data submission protocol for our database. With instructions covering our requirements, we should be able to collect spectral data with a certain quality much more quickly.

5 Digital data format and copyright

Question (Akira Ono)

I understood that all data were managed digitally in the development site while these were converted into analog format for the data open to the public so that the users from the Web were unable to access the digital data. Is this correctly understood?

The reason for a user not being able to access the digital data is because the spectral data acquired by AIST is copyrighted. When a third party requests to use the data, they have to pay a royalty. Is this correctly understood?

Answer (Takeshi Saito)

When the "digital data" in the question is "data consist of coordinate point information" and the "analog data" is a "GIF image data", then your indication is correct.

The main reason for the users not being able to reach the digital data format is not because of the rights or the royalties of copyright that we may receive but is for the protection of the copyright. This is based on the protection of SDBS from unjust

imitations and of others making profit by using the imitated spectral information of SDBS. The digital data have high workability so reproduced materials from the data have a high commercial value. If most of the digitally compiled spectral data were copied by others, they can construct a spectral database similar to SDBS or may be able to create more valuable database from the data. This is a menace for SDBS. We think even if the copyright infringement is recognized and can be appealed in court, we must spend considerable time and effort on the suit. Large amount of data copy is a copyright infringement even if the data is in GIF format and may cause SDBS to be in a threatening situation. Therefore, user access from the Web is monitored all the time.

When a third party wants to use data in the digital or the image formats, the licensing from AIST is necessary. In the case where the party wants to sell the data, they have to pay a royalty to AIST. We think this is a different point from the copyright in disclosing the data in the Web. We are not allowed to name names because of secrecy agreements, but we have several experiences of offering a large amount of data at a time, and in each case we have received payments of royalties based on the contracts. In the case of IR spectra, we have a contract to offer data with royalty when we release new spectral data. We also have experience of offering GIF image data to a database operated in the United States of America.

6 Comparison with other spectral databases
Question (Akira Ono)

I think there are spectral databases besides this one in the world, especially, those distributing the digital spectral information to private companies with a fee operated by private companies. Would you introduce such databases, and instruct us of the differences in roles and characteristics from this database?

Answer (Takeshi Saito)

There are not many spectral databases that are open through the Web. When looking at such free databases opened to the public, the number is very limited. We have not encountered a free accessible spectral database with this many ^{1}H NMR spectral entries with their spectral patterns and chemical shift assignments.

One of the free accessible databases on the Web is NIST Chemistry WebBook (http://webbook.nist.gov/) operated by the National Institute of Standards and Technology, NIST. This database compiles a variety of physical chemistry data as well as spectral data. This database is similar to RIO-DB of AIST because most of the data compiled in the WebBook are based on the research products of NIST. Besides 15,000 MS and 16,000 IR spectra, the WebBook compiles other spectra like an ultraviolet-visible absorption spectrum and a terahertz spectrum. There is no need to install software to browse the data of the WebBook. Spectra and other information are compiled along with the compound information, which you can search. We do not have a list of compounds compiled in the WebBook, however, it is our presumption that many generally used chemical reagents are compiled. This database works much like our database that serves as that of a public organization. MS data was sold as NIST 08 Mass Spectral Library which was released in 2008. Many

spectral data were obtained at National Institutes of Health (NIH) and United States Environmental Protection Agency (EPA) and were evaluated at NIST. The data can be purchased for use on an independent personal computer (PC). The number of data compiled in this database is about 220,000 spectra for 190,000 compounds, which is much more than those open through the Web. When we did collaborative research with NIST during our former Agency of Industrial Science and Technology era, many of our MS data were compiled in the NIST MS Library. These data must have been included in the NIST Mass Spectral Library. This data library can be searched with a fee with MS spectral patterns obtained by many kinds of MS spectral instruments.

SpecInfo releases data through the Web with access charge, and it consists of 90,000 ^{1}H NMR, 300,000 ^{13}C NMR, and many multinuclear NMR, IR and MS spectra. Data update does not seem to have been performed since 2006.

Looking at a domestic site, MassBank (http://www.massbank.jp/) which is a high resolution mass spectral database of metabolites is open free to the public. As of April 05, 2011, almost 31,000 spectra from 20 research institutes have been compiled. Target compounds are specified and limited to metabolites. Users of this database should download special software for this database, and install it on their computers. This software enables you to search, browse, display full and expanded spectrum, and register spectral data to the database. The database construction was initiated and has been supported by the project, "Integrated Database System for Metabolomics", under the Institute for Bioinformatics Research and Development of Japan Science and Technology Agency since 2006. As our database started with a research project and had developed into a solid base of our activity, we look forward to seeing how MassBank will be developed after this project is finished.

Databases from Bio-Rad are using a non web platform. The database covers data of SpecInfo, NIST MS Spectral Library in addition to Sadtler spectral data. Our NMR spectral data are also compiled. This database is sold as a package of data and software for handling the data, "KnowItAll", that needs to be installed on a computer. The compiled data are about 50,000 ^{1}H NMR, 430,000 ^{13}C NMR, 190,000 MS, 7000 Raman, and 230,000 IR spectra. Of our data, 13,000 and 11,000 ^{1}H and ^{13}C NMR spectra, respectively, are included. Users can use a spectral pattern for spectral search, search for mixtures, and more complicated searches can be performed with the software.

Sigma-Aldrich sells "Sigma-Aldrich Spectral Library" which has a total of more than 50,000 compounds of NMR, IR and Raman spectra. There are two formats of data on a computer and as a book.

When comparing a database operating on the Web and on an independent PC, there exist both advantages and disadvantages for each. For example, an advantage of the Web based database is instantaneousness. Our database can perform data addition and update easily; our database is updated twice every year to supply new data to our users. Many of the other databases introduced here do not seem to update their data that often. On the other hand, a database on the PC has better usability. For example, spectral pattern matching search can be done which many users find useful.