Recent Development of Nuclear Quadrupole Resonance Spectra Database *

Hideaki Chihara

Japan Association for International Chemical Information (JAICI)
Gakkai Center Bldg., 2-4-16 Yayoi, Bunkyoku 113, Tokyo, Japan

Koichi Mano

Department of Applied Chemistry, Faculty of Engineering, Osaka City University,
Sumiyoshi-ku, Osaka 558, Japan

Z. Naturforsch. **47a,** 446–452 (1992); received August 3, 1991

The Nuclear Quadrupole Resonance Spectra (NQRS) Database contains all the numerical data of resonance frequencies that have been published since 1950, the year NQR was discovered, through the end of 1989. The total number of records compiled is 10,155 as of May, 1991. One record usually covers data of a particular polymorphic modification of a crystalline substance. About 3/4 of all the records deal with organic or organometallic substances. Recent changes in the data content and format, development of online and stand-alone search software, statistics of the database, and updating problems are described. A personal computer version of the database with load/search software was prepared and is described in the Appendix.

I. Introduction

The database of nuclear quadrupole resonance spectral data which have been published in numerical form now contains 10,155 records. Building of the database began in 1980 by a team of Japanese scientists which was organized by the Japan Association for International Chemical Information (JAICI) to form the NQRS Data Committee with moral support of the International Symposium on Nuclear Quadrupole Resonance Spectroscopy (ISNQRS)[1]. Many researchers in the world helped the Committee by sending reprints and preprints of their research papers for inclusion in the database.

There have been some changes in the content of the database as techniques of data input advanced. The quantity of data increased to the extent that the database is almost up to date. Services include online, interactive search, dissemination in the form of magnetic tape and diskettes, and software for using the database on a personal computer. This paper describes such developments that occurred during the past decade.

II. New Data Structure

The database is a collection of logical records. One logical record in the case of the NQRS database consists of data related to a specific chemical substance. Since NQRS is sensitive to the state of aggregation of chemical species, different polymorphic forms usually make separate entries and therefore separate logical records unless there is a phase transition temperature between the forms.

The kinds of data contained in a record are typically the following:

1. Chemical Abstracts Service (CAS) Registry Number

The number of known chemical substances exceeded 10 million in 1990, and as a result the CAS Registry Number now takes 11 characters, e.g. 113780-43-1. The structure of the record was changed accordingly.

2. Molecular Formula

Molecular formula (field FORM) was expressed by capital letters only, e.g. C5CL15HG1 but a new data element (a field of the database), MOLF, has been created to include formulas using lower-case letters, C5C115Hg. This is easier to read when printed with

* Presented at the XIth International Symposium on Nuclear Quadrupole Resonance Spectroscopy, King's College London, July 15–19, 1991.

Requests for reprints and details of the personal computer version should be addressed to Dr. H. Chihara, Japan Association for International Chemical Information (JAICI), Gakkai Center Bldg., 2-4-16 Yayoi, Bunkyo-ku, Tokyo 113, Japan (fax: +81-3-3816-7826).

0932-0784 / 92 / 0100-0446 $ 01.30/0. – Please order a reprint rather than making your own copy.

numbers as subscripts. However, the previous format is also useful in sorting the substances according to the formula and therefore the FORM field is retained for the purpose of database management. In the distribution file, the data in the FORM field are replaced with the data in the MOLF field, thus providing only the C5C115Hg type of formula. Old data were converted to the new form by a program.

There is a question of how to handle substances of non-stoichiometric composition. This question has become progressively important as NQRS data are increasingly reported for ceramic materials which become superconductive at high temperatures.

3. Chemical Names

Various chemical names are provided, systematic CAS and/or IUPAC nomenclature, common names, trade names, etc., when they appear in the original documents. These names were written all in capital letters in the original version of the database, but now lower-case letters are also used. Capital-letter data are generated from the new data by a program and retained in the archival file of the database. Inversely the old data that used only capital letters were converted to the new form also by a program. Symbols [and] have also been introduced.

```
NUMB    *022                                      Q012203
MOLF         74-05-8 C2H3N1                       Q012203
FORM         75-05-8 C2H3N                        Q012203
SUBS    ACETONITRILE;                             Q012203
NAME    Acetontrile;                              Q012203
MODF    .BETA. phase;                             Q012203
FREQ    P14    77. 2.8111   2.7952               1Q012203
FREQ    P14   148. 2.7915   2.7821               1Q012203
FREQ    P14   215. 2.7586   2.7565               1Q012203
FREQ    P14   225. 2.7512   2.7496               1Q012203
ATHR    GOURDJI  M; GUIBE L;
        KAPLAN A; PENEAU A;                      1Q012203
AUTH    Gourdji M; Guibe L;
        Kaplan A; Peneau A;                      1Q012203
JRNL    JMOSB4 J Mol Struct; 111; ; 1983;
        371; 377;                                1Q012203
ATHR    NEGITA H; CASABELLA P A;
        BRAY P J;                                2Q012203
AUTH    Negita H; Casabella P A;
        Bray P J;                                2Q012203
JRNL    JCPSA6 J Chem Phys; 32; ;
        1960; 314; 315;                          2Q012203
KYWD    QCC; ETA; POLYMORPHISM;
        TEMP DEPENDENCE;                          Q012203
REMK 1  This phase exists down to 210 K,
        and between liq. N?2? temp. and 6        1Q012203
REMK 2  OK; It transforms to .ALPHA. phase
        on annealing at 210K;                    1Q012203
REMK 1  (QCC, ETA) = (3.7375 MHz, 0.82)
        for N-14; .beta. Form is obtained by
        rapid quenching of sample to 77 K;       2Q012203
REMK 2                                           2Q012203
END                                               Q012203
```

Fig. 1. Typical record of NQRS database.

4. Polymorphic Modification

Crystal modifications are given in the MODF field as they are given in the original document.

5. Frequency Data

Besides nuclear mass number, temperature of measurement and resonance frequencies, the method of measurement has been added: C for continuous method, P for pulse method, M for NMR, X for unknown method. There is a disadvantage in representing a nuclear species in terms of a nuclear mass number, i.e. ^{64}Zn and ^{64}Ni cannot be distinguished merely by the mass number. While this is theoretically possible, it has never happened (^{64}Zn is a shortlived radioactive element and ^{64}Ni has an abundance of only 1.16%; both species do not have a quadrupole moment!).

Only published numerical data are included in the database. Those data that are reported only in the form of graphs are not taken into consideration at present. However, if quadrupole coupling values are reported with or without asymmetry parameter values, the fictitious resonance frequencies are calculated from them, in which case the REMK field is used to note the fact.

6. Author Names

As for chemical names, author names are now recorded both in capital letters only and in capital/lower-case letters. Old data in capital letters (archived in ATHR field) only were converted to the latter form by a program. However, the conversion is difficult algorithmically and a small number of author names in the older records may not have been correctly converted.

7. Journal Data

The field JRNL contains, as in the original version, CODEN, journal name, volume, issue, year of publication and first and last pages.

8. Keywords and Keyphrases

There has been no change in the principles of selecting keywords but as science advances new keywords are being added.

9. Remarks

The REMK field allows the data analysts to write any additional information in natural language: They are encouraged to give specifically the method of preparing the specimen studied, the stability of the substance, and the use of special solvents like liquid crystals. Also quadrupole coupling constant and/or asymmetry parameters are recorded in this field. The format reporting these data has been standardized as in Figure 1.

A record thus consists of 11 fields, FORM, MOLF, NAME, SUBS, MODF, FREQ, AUTH, ATHR, JRNL, KYWD, and REMK. The fields MOLF, SUBS, and ATHR, are contained only in the archival file. The rest is contained in the distribution file. The distribution file is in the card image format with 80 colums. The first card or line shows the total number of lines of that record including the last END line.

III. Database Updates

The archival file is continuously updated as new data appear in the literature whereas the distribution file is updated twice a year.

Literature is being surveyed regularly. The systematic survey is done by searching the CA File online on the STN International service by use of a specially developed search profile and original documents are collected which correspond to the relevant hits of the search. Securing the original documents is extremely important for accuracy and relevancy of the database but at the same time it is the hardest part of the work. When document delivery service of the university library network fails to supply us with a document, we ask the author for a reprint. Authors have always been cooperative in this regard and often send us additional information.

Original documents and data preparation sheets are sent out to members of the NQRS Data Committee for extracting and evaluating information. For the sake of uniformity of record content and quality, a detailed input manual is available. Evaluation of data is not always possible because a recent trend is that short communications in the form of Letters to the Editor are more common, which do not describe experimental details. Errors in frequency values (and quadrupole coupling constants) are not usually reported. In quite a few cases, a new paper deals with a substance for which NQRS data have already been published. This point will be treated in a later Section.

The data preparation sheets having been filled out, they are returned to the editorial office at Osaka University where the data are checked for consistency and input grammar. Some data are missing at times, e.g. the CAS Registry Number and CODEN of the journal, which are not easy to find. Keywords and keyphrases are added and remark notes are edited. Such checking is done on a personal computer by use of a series of programs as far as being practical. What is the most time-consuming part in the editorial process is duplicate detection/elimination and merging data on the same chemical substance.

IV. Some Statistical Figures

Here are some numbers relating to the whole database. These numbers reflect the research activities of this science worldwide.

1. Total number of records is 10,155 and the total number of lines is 153,044; therefore an average record has 15.07 lines (the minimum requirement is seven lines).

2. The largest record, #11434 for 1,3,5-Trinitrohexahydro-*s*-triazine ($C_3H_6N_6O_6$), has a maximum of 70 lines.

3. There are 9,140 unique CAS Registry Numbers, but there are 471 records which lack Registry Number information. This means that 10% of substances have separate records because of the existence of polymorphic modifications.

4. The number of inorganic substances in the database is 2,777 and the number of organic substances is 7,378. The organic class includes organometallic compounds.

5. On the average, one record contains 1.57 chemical names. The record #3199 for 2-Thiouracil ($C_4H_4N_2OS$) has a maximum of 8 synonyma.

6. 74 different nuclear species have been studies from ^{2}D to ^{209}X. The ten most frequently studied nuclei are ^{35}Cl (5,307 substances), ^{14}N, ^{79}Br, ^{127}I, ^{81}Br, ^{2}D, ^{121}Sb, ^{123}Sb, ^{75}As, and ^{59} Co in this order.

7. There are 279 sources from which data were extracted. These include scientific journals, monographs, data compilations, conference proceedings, dissertation etc. The most frequently cited journal is J. Chem. Phys. from which 1,402 data were extracted. Others are

Izv. Akad. Nauk SSSR Ser. Khimiya (1,018); J. Magn. Resonance (625); Doklady Akad. Nauk SSSR (497); Inorganic Chemistry (469); J. Mol. Structure (441).

8. The total number of keywords is 40,740 including duplication. The number of keyphrases is 24,963. When a keyphrase, e.g. relaxation time, is adopted, it is broken down into words and "relaxation" and "time" are also taken as keywords.

9. The number of records which have REMK data is 5,125.

10. The number of records which have MODF data is 563.

11. The distribution of records over publication years is as follows.

Year	No. of records	Year	No. of records
1950	7	1970	386
1951	28	1971	668
1952	48	1972	478
1953	95	1973	453
1954	122	1974	625
1955	62	1975	1393
1956	95	1976	596
1957	70	1977	492
1958	85	1978	402
1959	89	1979	551
1960	216	1980	522
1961	83	1981	103
1962	103	1982	468
1963	90	1983	524
1964	150	1984	330
1965	191	1985	444
1966	329	1986	465
1967	444	1987	443
1968	489	1988	169
1969	432	1989	129

V. Online and Other Dissemination Services

The database was loaded as an online file at the Computer Center of Osaka University and has been in service to Japanese academic institutes since 1981. The search keys chosen were

CAS Registry Number; Molecular Formula; Chemical Name; Frequency Range (with or without mass number specified); Author Name; CODEN of journal; Publication Year; Keyword or Keyphrase and Record Number.

The searcher may use Boolean logic operators, OR, AND, and NOT between the answer sets. There are 8 different display formats to choose from.

Due to regulations from the Government, this online service cannot be made available outside the Japanese academic community. Therefore, the database has been licensed in the form of magnetic tapes or diskettes.

A recently developed small software system facilitates loading of the database and search on a personal computer. The system called PC SEARCH uses Turbo C language, runs on MS-DOS and has a search/display capability almost identical to the online service from Osaka University. Outline of the structure of the PC SEARCH is given in the Appendix.

VI. Duplication of Work

While building and updating the NQRS database, we encountered numerous cases in which research work was conducted apparently without knowledge about previously published results on the same substance. There also are papers published by the same author (or almost the same group of authors) in different journals, which deal with the same substance. Among these, some report improved measurements but the majority reports the same results as if it were a new investigation (no references to previous papers).

Such duplication is detected and comes to our attention only after we make cross-checking by use of the CAS Registry Number. The following figures are the results of updating operation during the past year.

Total number of records at the end of June, 1990 ...	9147
Number of data sheets prepared during the year ...	1403
Total number of records at the end of June, 1991 ...	10155
Net increase in the number of records (new subs.)...	1008
Number of revised records	74

Therefore, 321 out of 1403 input, or 23%, were purely duplicated reports.

To avoid wasting resources, the NQRS Database is useful and can help identifying the previously published data by inputting the CAS Registry Number, Molecular formula, or chemical names before making measurements or even before planning an investigation.

Acknowledgement

The author (H. C.) expresses his sincere gratitude to the fellow scientists, especially those who contributed to building the database; Dr. T. Asaji, Dr. Y. Furukawa, Professor M. Hashimoto, Dr. K. Mano, Professor D. Nakamura, Professor N. Nakamura, Professor T. Okuda, and Professor H. Terao. Professor N. Nakamura worked as a key person for the routine operations of the Committee.

Appendix

A Personal Computer System for Retrieval of NQRS Data

Specification of the System

Software was developed which allows the NQRSDB to be loaded and used on a personal computer by a researcher who has no prior knowledge of programming, database management and online searching. The system then has to have a maximum flexibility regarding the hardware requirements and must run in a standalone fashion, i.e. independent of the need for other software as far as possible. The specifications were thus the following.

1. Target PC: IBM's PC/XT, AT, PS/2, NEC's PC9801 series.
2. Programming Language: Turbo C Version 2.0. Only ANSI C standard functions may be used.
3. Any commercial software other than MS-DOS and Turbo C may not be used. If the PC edition (NQRSPC) were made to be based on such other software, users would also have to purchase that and the portability of the system would then be restricted.
4. NQRSPC thus consists of the following three modules.
   a) MAKE_INDEX_FILES: The preprocessors extracting the index files from the DB in advance.
   b) PC_SEARCH: The main module including search program.
   c) SETUP: Installation program.

Database and Searchable Elements

The following data elements were chosen as the searchable items; CAS Registry Number, Molecular Formula, Substance Name, NQR Frequency, Author Name, CODEN, Publication Year, and Keyword.

Method

1. Outline

In order to reduce the searching load, the 8 index files (IDF) corresponding to the search items are extracted from the DB by the respective preprocessors. The IDF stores the search items (e.g. Author Name) and their corresponding record numbers (RNs) which point to the address of the items on the DB. Searching begins by entering a search key (e.g. Smith, J. A. S.). When the system receives the query key from the keyboard, it sets a search number (SET No) at first. Then the system begins to search the key term on the IDF (Author index file in this case) and writes the RNs of the hit item(s) into the user's work file. The RNs thus obtained are sorted in the ascending order and duplicating numbers are eliminated. The set of such RNs is stored in the array whose top address is pointed at by the pointer which is referred to as SET No. The system finally displays the total number of hits on the screen.

2. Index Files

The searching time depends on the search algorithm and the IDF format. We decided on the search method to be used, judging from the property of the search element and the number of the relevant RNs. The structure of the index file will be explained briefly.

With regard to the Substance Name Index, the record length of the file is normalized to that of the longest name. Three kinds of truncations, i.e. right-hand, left-hand, and both-hand truncations) were made possible in a query for the **Substance Name.** When users access to this IDF, most of them will try both hands truncation. In such a case, the total number of hits can be very large. Therefore, it is proper that the IDF is sorted in the ascending order of RN and linear searching is adopted as the algorithm because the system can save time for sorting such a large number of RNs after having searched the key name. Indexes of CODEN and YEAR were made in the same way.

Binary search was adopted as the algorithm of searching the Registry Number and Molecular Formula elements.

There may be many RNs corresponding to one author name. This element of the IDF is sorted in two directions, that is, the **Author Names** are sorted in alphabetical order (vertically) and the RNs are sorted in the horizontal direction for each **Author Name.** The top position of each **Author Name** is found using binary

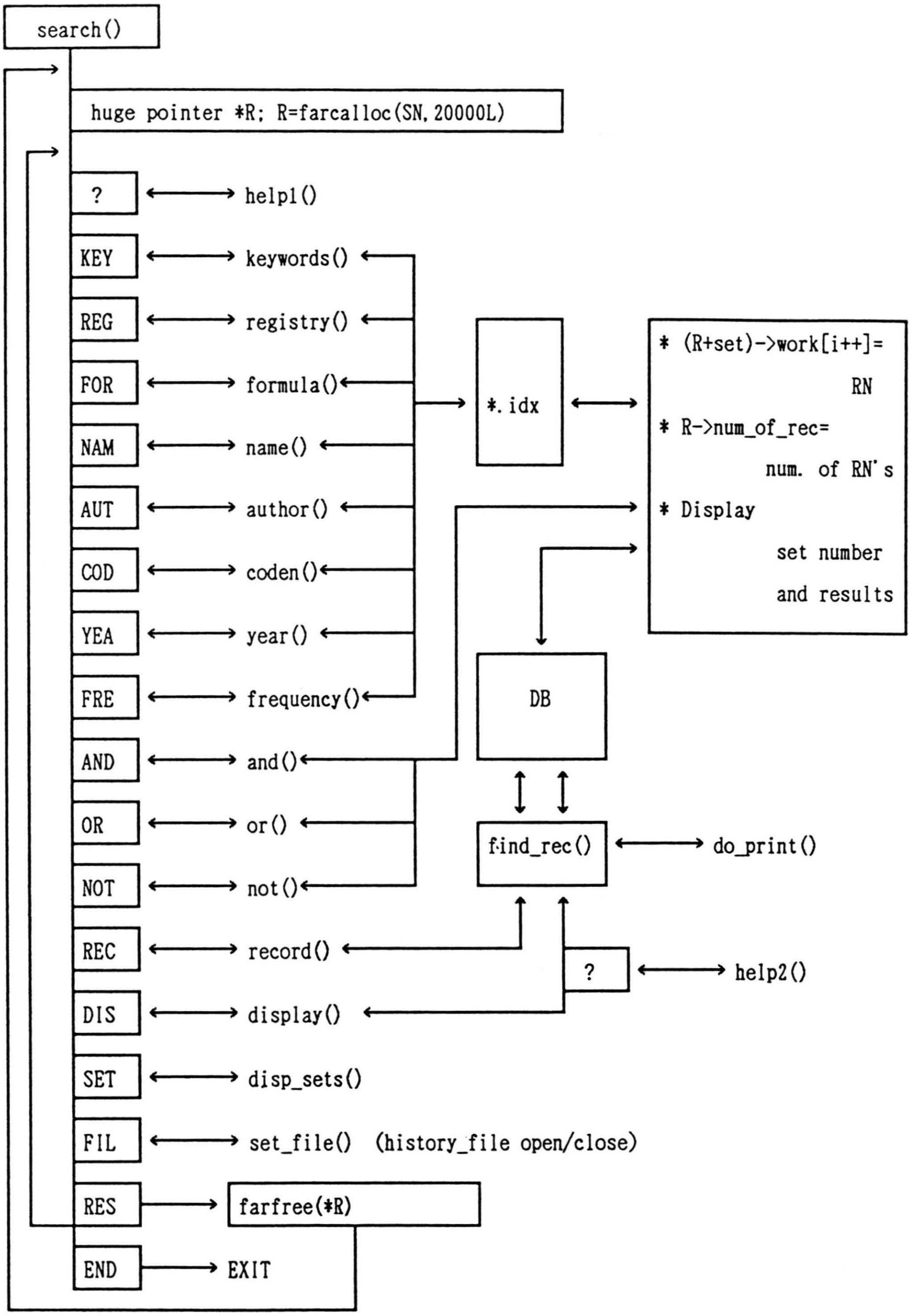

Fig. 2. Flow chart of search function.

search method and **fseek** function. The structure of the Author Index is thus

```
ABDULLIN R S;
  001142 002504 003421 005814 006167 006731
ABDULLIN R S;
  011625 011762 011763 011764 011765 011766
ABE Y;
  000528 003363 003608 005631 005632 008356 008708
```

In the case of frequency data, users must specify the nuclear mass number (N) and the range of frequencies (the lowest one, F1 and the heighest, F2). The searching proceeds as follows: (i) The system first finds the top address (by N) of the memory region at which the target nuclear data are stored. (ii) It then moves the file pointer to that address. (iii) The system judges whether the datum lies between F1 and F2 or not by comparing the top of the frequency data with F1 and/or F2. (iv) If the judgment was negative, it continues to compare F1 and/or F2 with the frequency data which have been sorted in the descending order one by one until getting a positive answer in the judgment. To do the above procedure efficiently, a special format of index file was devised. The first 11 lines of the IDF record all the top addresses of memory blocks, each storing data of a nuclear species like in the following:

```
2;11  7;617  8;673  9;675  10;679  11;695.
```

This means that the frequency data of D (mass number 2) starts at the $(11+1)$-th line.

3. *Search Functions*

Standard capabilities of search software have been incorporated in the PC edition, i.e. Boolean logic operators between hit sets, truncations, opening/closing of a history file, pre-formated display options etc. The flow-chart of the search function is illustrated in Figure 2.

Loading

The whole **PC_SEARCH** system (size of ca. 14 Mb) was compressed and changed to the auto-decompressing file, **NQRS. EXE**, by using an archiver program, **LHarc. EXE** [2], **NQRS. EXE** file (size of ca. 2.5 Mb) was stored in 4 sheets of 2 DD mini-floppy disks using a DOS command, **BACKUP.** SETUP command creates a directory, ⟨**NQR**⟩, in the hard disk drive specified by the user, re-stores **NQRS. EXE** in the ⟨**NQR**⟩, and executes **NQRS. EXE** to expand all the files into the **PC_SEARCH** system of the original size.

Implementation

PC_SEARCH ran successfully on NEC's PC9801/RA21, IBM's PS/5530Z, PS/55 note, PS/2, and several IBM compatible machines. The searching time is typically within 15 seconds under the condition of NEC's PC9801/RA21 and ICM HC130 (130 Mb hard disk, average seek time is 19 ms).

[1] H. Chihara, J. Mol. Struct. **83,** 1–7 (1982).
[2] H. Yoshizaki, C Magazine **3** (1), 59–68 (1991).

Nachdruck — auch auszugsweise — nur mit schriftlicher Genehmigung des Verlages gestattet
Verantwortlich für den Inhalt: A. KLEMM
Satz und Druck: Konrad Triltsch, Würzburg