DOCUMENT RESUME

ED 048 906 LI 002 715

AUTHOR	Schipma, Peter B.; And Others
TITLE	Comparison of Document Data Bases
INSTITUTION	Illinois Inst. of Tech., Chicago. Research Inst.
PUB DATE	Jul 70
NOTE	36p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Abstracting, Automatic Indexing, Bibliographic Citations, Biology, Chemistry, *Data Bases, Engineering, *Information Processing, *Information Retrieval, Information Storage, *Magnetic Tapes
IDENTIFIERS	*Machine Readable Bibliographic Data Bases

ABSTRACT

This paper presents a detailed analysis of the content and format of seven machine-readable bibliographic data bases: Chemical Abstracts Service Condensates, Chemical and Biological Activities, and Polymer Science and Technology, Biosciences Information Service's BA Previews including Biological Abstracts and BioResearch Index, Institute for Science Information Source Tape, and Engineering Index COMPENDEX. Selected issue test tapes of each data base were printed and checked for the types of data that were contained in the issue and the methods in which the data were formatted. This paper compares the physical formats of the tapes and describes the varied treatments given to such data elements as authors, titles, abstracts, etc. Great discrepancies in the presentation of essentially similar bibliographic data were found, and some suggestions for mitigating the discrepancies by use of standards are offered. (Author)

COMPARISON OF DOCUMENT DATA BASES

Prepared by

Peter B. Schipma
Research Scientist, Information Sciences

Martha E. Williams
Manager, Information Sciences

Allan L. Shafton
Technician, Information Sciences

July, 1970

ABSTRACT

This paper presents a detailed analysis of the content and format of seven machine-readable bibliographic data bases: Chemical Abstracts Service Condensates, Chemical and Biological Activities, and Polymer Science and Technology, Biosciences Information Service's BA Previews including Biological Abstracts and BioResearch Index, Institute for Science Information Source Tape, and Engineering Index COMPENDEX.

Selected issue test tapes of each data base were printed and checked for the types of data that were contained in the issue and the methods in which the data were formatted. This paper compares the physical formats of the tapes and describes the varied treatments given to such data elements as authors titles, abstracts, etc. Comparison of data bases requires common use of terms. All terms are defined at the beginning of the paper.

The authors found great discrepancies in the presentation of essentially similar bibliographic data, and they offer some suggestions for mitigating the discrepancies by use of standards.

COMPARISON OF DOCUMENT DATA BASES*

by

Peter B. Schipma**
Martha E. Williams
Allan L. Shafton

IIT Research Institute

INTRODUCTION

The Information Sciences section of IIT Research Institute (IITRI) operates the Computer Search Center (CSC), which provides retrieval of information from machine-readable data bases for scientists in industrial and academic organizations. One of the goals of IITRI in establishing the Center was to provide search service from a variety of data bases. At this writing, three data bases are used, and an operational capability exists for four more. In attaining this capability, Center personnel made a detailed evaluation of several data bases, and a summary comparison of them is presented herewith. We would like to note that there are variances from issue to issue of any data base, and that our evaluations were based on the following issues:

---

*Financial support for this work was provided by the Chemical Information Unit of the National Science Foundation.

**Please send galleys to Peter B. Schipma.

Chemical Abstracts Service	Chemical and Biological Activities (CBAC) vol. 8, issue 13, 1969
Chemical Abstracts Service	Polymer Science and Technology (POST) vol. 4, issue 2, 1969
Chemical Abstracts Service	CONDENSATES vol. 70, 71, 72 issues 1-26, 1969
Biosciences Information Service	Biological Abstracts vol. 51 issue 6, 1970
Biosciences Information Service	BIOResearch Index vol. 51, issue 3, 1970
Engineering Index	COMPENDEX Test Tape, April, 1969
Institute for Scientific Information	ISI Source Test Tape, 1969

Part I - DEFINITIONS

The following definitions are given in order to resolve some of the ambiguities that can arise from the differing terminologies used by the various tape suppliers, manufacturers and information processors. The definitions are of two types--conceptual and physical. The definitions for logical record, component, element, etc. are conceptual or content-oriented, i.e., they deal with the information content that is in a data base and is to be made available to users. The terms block, physical record, field, etc. are physical terms that relate to the arrangement of information on a tape, and are

of interest from a processing point of view. These definitions should be referred to for interpretation of this paper.

Data Base

A data base is a machine-readable group of like items. For the purpose of this paper each data base is on tape and is issued in segments according to a fixed schedule.

Item

An item is that which is identified by an accession number--it may be a book, document, journal article, document surrogate, symposia proceedings, conference paper, or any other entity that is uniquely identified by an accession number.

Logical Record or Unit

A logical record or unit is that portion of a data file that contains all information and data pertaining to one item. For example, all information related to one CAS Condensates citation (accession no., volume no., issue, CODEN, primary source, date, title, author(s), abbreviated journal title, research site, and index terms) is included in one logical record, along with all tags, identifiers, and other descriptive information, which will be discussed later.

Component

A component is a portion of a logical record made up of one or more elements related to a single type of information. For example, in CAS Condensates, one component of a Condensate record contains the following elements: accession number, volume number, issue number, data type designator, and data; whereas in ISI's source, each component is a single element whose position in a fixed field implies content information which might otherwise be specified by additional elements.

Element

An element is the smallest unit of data. It may be all or a portion of a component depending on the tape service involved. In CAS Condensates, several elements make up a component--whereas in ISI's source, elements are identical with components, i.e., content information is identified by virtue of its position in a fixed field. Elements may be data or tags to data. A data element is a type of element that conveys information directly related to the item. A tag element is a type of element that conveys information about the item or about specific data elements within an item.

Examples

Data Elements	Tag Elements
Smith, John	03 - a numeric code indicating that the data attached is an author name (used in CAS Condensates)
J. Chem. Doc.	
JCHDA	126 - a number indicating the length of a record (used in ISI Source Tapes)

Data element information is content information that is available for the users whereas tag element information is information that is descriptive and is of use to the processor.

Block

A block relates to a physical file and is that portion of a file that is transferable in one operation from an external storage medium to the central processor.

Physical Record

A physical record is a fraction of a file--it is a unit that is processed as a discrete entity. The number of bits that make up a record is determined by the tape supplier and varies from tape service to tape service, e.g., a Condensate physical record has a maximum of 8000 bits (1000 bytes) while BA Previews has a maximum of 12000 bits (500 bytes). One or more physical records comprise a BLOCK.

Format

The term format is used in two senses - one referring to the arrangement of records in a file, in terms of fixed or varying length of record, and the other referring to physical arrangement of elements and/or components within a physical record in a file.

Field

A field is a portion of a physical record assigned to an element. It can be of fixed or varying length. One or more physical fields comprise a physical record.

Part II - PHYSICAL ORGANIZATION

The first parameter of each of the data bases which we would like to delineate is that of representation of a logical record. This varies from data base to data base, some using one physical record to represent a logical record, and others using several physical records per logical record. All descriptions of physical records are based upon character representation, one character per byte. Such description is possible since all tapes studied were received in 800 bpi density, 9-track, EDCDIC mode, i.e., essentially IBM 360 mode.

LOGICAL RECORD

Condensates

The CAS Condensates data base is issued on a weekly schedule, with two tape issues corresponding to one printed issue of Chemical Abstracts, which the tapes represent. Twenty-six weekly tapes represent one volume. Volumes are numbered consecutively, and issues within each volume are numbered from 1 to 26. Tapes are available with or without IBM standard labels, and if labels are used they contain a tape serial number of 000000, and a tape dataset name of CAISSV. Each physical record on the tape is of varying length (maximum of 1000 bytes) and they are blocked to a size of 1200 bytes. Each physical record represents one component of a logical record. Thus several physical records are used to represent one logical record. Each physical record contains the following elements:

Bytes	1-7	CA abstract number
Bytes	9-10	CA volume number
Bytes	11-12	CA issue number

Byte 14 CA code for data type

Bytes 17-n Data

Bytes 8, 13, 15 and 16 are blank. There are variations in the number and content of physical records that comprise one logical record. The CA Codes for data type and the types of data for which they stand are:

1 CODEN and bibliographic information

2 Title of item

3 Authors and/or editors of item

4 Short journal title (country in the case of patents) and research site (author's full name and company of assignment in the case of patents). Both components are represented with a code of 4 and are distinguishable by the order in which they appear

5 Index terms and/or phrases (keywords)

The total number of physical records comprising one logical record is variable, since one physical record is used for each type 5 component, of which there may be none or several. Some physical records may contain all elements

except data, in which case it is indicated that data of the type given was not available. The usual order of records by type code is 1,4,2,3,4,5,...,5. The first appearance of a type 4 record for one logical record denotes that the component is short journal title, and the second appearance of a type 4 record indicates that the component is research site. The order of appearance of the type 3 record and the second type 4 record is reversed in the case of patents.

POST J

The CA Polymer Science and Technology (for Journals) data base is issued on a bi-weekly schedule, with 13 issues comprising one volume. Volumes are numbered consecutively and issues within a volume are numbered from 1 to 13. Tapes are available with or without IBM standard labels.

Each physical record on the tape is of varying length (maximum of 1000 bytes); records are blocked to a size of 1200 bytes. Each physical record represents one component of a logical record, and

contains the following elements:

Bytes	1-5	Abstract number
Bytes	6-7	Abstract number modifier
Bytes	9-10	Volume
Bytes	11-12	Issue
Byte	14	Code for data type
Bytes	17-n	Data

Bytes 8, 13, 15 and 16 are blank. There are variations in the number of physical records that comprise one logical record. The codes for data type and the types of data for which they stand are:

1 CODEN and bibliographic information

2 Title of item

3 Authors and/or editors of item

4 Short journal title and research site, distinguished by order of appearance of records

5 Keywords and compound name (CA preferred name), distinguished by order of appearance

6 CA Registry Number

7 Molecular formula

The total number of physical records comprising one logical record is variable, since one physical record is used for each type 5, type 6, and type 7 component, of which there may be none or several. A physical record may contain all elements except data, indicating no data for that type was available. The order of physical records by data type is 1,4,2,3,4,5,6,7,5,5,6,7,5...5,6,7,5,...5. The first record with type code of 4 is the short journal title, and the second is the research site. If a physical record with a code of 5 is preceded by a physical record with a code of 4 or 5, then it is a keyword record; but if it is preceded by a physical record with a code of 7, then it is a compound name record. A physical record with a code of 6 is always followed by records with codes of 7 and 5 respectively. These three records give the Registry Number, molecular formula and compound name of a chemical mentioned in the item being recorded. Thus if there are 6 keywords in a logical record and 3 chemicals are to be referenced, the physical records would appear as:

CODEN
Short journal Title
Title of item
Author(s) of item
Research site

Keyword 1

Registry Number 1

Molecular formula 1

Compound name 1

Keyword 2

Registry Number 2

Molecular formula 2

Compound name 2

Keyword 3

Registry Number 3

Molecular formula 3

Compound name 3

Keyword 4

Keyword 5

Keyword 6

CBAC

The CAS Chemical and Biological Activities data base is issued on a bi-weekly schedule, with 13 issues comprising one volume. Volumes are numbered consecutively and issues within a volume are numbered from 1 to 13. Tapes are available with or without IBM standard labels.

Each physical record on the tape is of varying length (maximum of 4000 bytes); records are blocked to a size of 4000 bytes. Each physical record represents one component of a logical record and contains the

following elements:

Bytes	1-5	Abstract number
Bytes	6-7	Abstract number modifier
Bytes	9-10	Volume number
Bytes	11-12	Issue number
Byte	14	Code for data type
Bytes	17 or 18-n	Data

Bytes 8, 13, 15, 16, and sometimes 17 are blank. The codes for data types and the types of data for which they stand are:

1 CODEN and bibliographic information

2 Title of item

3 Author(s) of and/or editor(s) of item

4 Research site

5 One sentence or meaningful phrase of the digest

6 CA Registry Number

7 Molecular formula

The total number of physical records comprising one logical record is variable, since the digest (notation of content) is broken up into phrases and/or sentences, each of which is recorded in a separate physical record with a type code of 5. The normal order of physical records by code type is 1,2,3,4,5...5,6,7,5...5. Several

sentences or phrases of the digest are contained in consecutive records until a chemical is encountered, at which point Registry Number and molecular formula records appear, to be followed by more of the digest. Some of the digest records have the data beginning in byte 18 rather than 17, according to no immediately discernible pattern.

BA Previews

The BA Previews data base is issued in two series, Biological Abstracts on a twice-monthly basis and BioResearch Index monthly. Twenty-four issues of BA comprise a volume as do twelve issues of BioRI. Volumes are numbered consecutively from 1 to 24 for BA and from 1 to 12 for BioRI. Tapes are available only without labels.

Each physical record on the tape is of varying length (maximum of 1500 bytes) and they are blocked to a size of 3600 bytes. Each physical record represents one component of a logical record, and contains the

following elements:

Bytes	1-2	Volume number
Bytes	3-8	Abstract number
Byte	9	Code for data type
Byte	10-n	Data

Six physical records comprise one logical record. The codes used by BA for data types and the types for which they stand are:

1 CODEN and abbreviated journal title

2 Bibliographic information

3 Author(s) and/or editor(s) of item

4 Title of item (augmented in most cases)

5 CROSS index codes

6 Biosystematic index codes

The physical records appear in the above order. The CROSS index and Biosystematic index codes are five digit numbers appearing in six-character fields. The sixth position may contain a blank, an asterisk or a hyphen. Several of these code numbers usually are present in a physical record of type 5 or 6, in consecutive six-character fields.

ISI Source

The Institute for Scientific Information Source Tape data base is issued on a weekly schedule, with 52 issues comprising one volume. Volumes are numbered consecutively and issues within a volume are numbered from 1 to 52. The tapes are available with no labels only.

Each physical record on the tape is of varying length, with a maximum size of 246 bytes. The records are unblocked. There are three types of physical records, named for convenience:

Primary Author

Trailer Primary Author

Secondary Author

The elements contained in each of these types of physical records are given below:

Primary Author Record

Bytes	1-5	Length of record
Byte	6	Constant "K"
Bytes	7-13	Seven-digit source article number

Bytes	14-24	Primary source
Bytes	25-35	Blank (ϕ)
Bytes	36-46	Source journal
Bytes	47-50	Source volume
Bytes	51-54	Source page
Byte	55	Trailer record sort key (ϕ if no trailer to follow; non-blank if trailer to follow)
Bytes	56-57	Source year
Byte	58	Code indicating type of source item
Bytes	59-61	Number of references
Bytes	62-63	Number of secondary author characters
Bytes	64-69	Number, supplement, part
Bytes	70-74	ISI journal issue accession number
Bytes	75-245	Secondary author (up to 9); followed by the article title data

Byte	246	Record mark. If the title data are completed before column 246, the record is shortened and the record mark set in the lowest position divisible by 6 that the data will allow

Trailer Primary Author Records

Bytes	1-5	Length of record
Bytes	6-54	Identical to position 6-54 in the first primary record
Byte	55	"2"-"9". "9" is the last regardless of previous number in position 55
Bytes	56-58	Identical to position 56-58 of first primary record
Bytes	59-63	⅋
Bytes	64-74	Identical to position 64-74 of first primary record
Bytes	75-245	Title overflow
Byte	246	Record mark. The record mark is again set in the lowest position divisible by 6

Secondary Author Records

Bytes 1- 5	Length of record
Bytes 6-13	Identical to position 6-13 of first primary record
Bytes 14-24	Secondary author
Bytes 25-35	Primary author
Bytes 36-54	Identical to position 36-54 of first primary record
Byte 55	"S"
Bytes 56-58	Identical to position 56-58 of first primary record
Bytes 59-61	⊅
Bytes 62-63	"00"
Bytes 64-65	⊅
Byte 66	Record mark

Each logical record comprises a variable number of physical records in accordance with the number of authors of the item. For the primary author, there is one physical record which contains full information about the source, title, etc. If the title is too large to fit in the space

allotted, trailer primary author records are used for the overflow. If, on the other hand, the title is shorter than the maximum useable space, the record is shortened to the nearest byte value divisible by six. In addition to the primary author record (and trailers, if required) there is one secondary author physical record for each of the other authors of the item. The physical records comprising one logical record appear in the following order:

secondary,...,secondary, trailer,...,trailer, primary.

COMPENDEX

The Engineering Index COMPENDEX data base is issued on a monthly schedule, with 12 issues comprising one volume. Volumes are numbered consecutively, and issues within a volume are numbered from 1 to 12. Tapes are available with no labels.

Each physical record is of varying length (maximum of 8000 bytes; records are blocked to a size of 8000 bytes. All information for one logical record is contained in one physical record. The record begins with:

Bytes	1-2	"EI"
Bytes	4-5	Volume number
Bytes	7-8	Issue number
Bytes	10-13	Internal sequential identification number

The remainder of the physical record contains an abstract number, authors, abbreviated journal title, bibliographic information, title, and abstract, all of which are of varying length and denoted by code numbers between the data fields.

Part III - CONCEPTUAL ORGANIZATION

Having defined the composition of a logical record in terms of the physical record(s) used for each data base in the preceding section, in this section we compare the contents of the data bases in terms of the components they contain and the information contained in the components.

Condensates, POST J, CBAC and COMPENDEX contain the volume and issue numbers in each physical record. BA and BioRI contain only the volume number, the issue being determined by

reading the printed label on the tape reel. ISI gives volume and issue numbers in the tape label.

Accession Number

Accession numbers are numbers that uniquely identify items within a file. With respect to number of characters, they range from 5 to 7 digits, and they vary with respect to presence of a check character.

Condensates	7 characters 6 digits (from 000001 to NNNNNN for each volume) with a 7th check character
POST J	5 characters 5 digits No check character
CBAC	5 characters 5 digits No check character
BA Previews	5 characters 5 digits No check character
ISI Source	7 characters 6 digits 7th character A thru I or Ø

COMPENDEX — 5 characters

5 digits

No check character

Author

Condensates — Authors' names are given as follows:

SMITHØJS,ØJONESØHL,ØJOHNSONØPU,.

Condensates tapes include the total number of authors unless the number of characters exceeds 1000.

POST J — Same as Condensates

CBAC — Authors' names are given as follows:

SMITHØJS,ØJONESØHL,ØJOHNSONØPU.

CBAC author format is the same as that of Condensates and POST except that the last initial of the last author's name is followed by a period instead of a comma and a period.

BA Previews — Authors' names are given as follows:

SMITHŁJØS/JONESŸHØL/JOHNSONŻPØU

All authors are given unless total number of characters exceeds 1500.

ISI Source

Authors' names are represented as SMITHbJbbb, always in an 11 character field. If the name is less than 11 characters it is padded on the right with blanks. If it exceeds 11 characters it is truncated on the right until only one initial is left. If it then still exceeds 11 characters, the surname is truncated from the right to eight characters and a period added. Thus "HOFMEYERbJRS" would be truncated to "HOFMEYERbJR", and "RUMPELMEYERbPR" would be truncated to "RUMPELME.bP". As noted in the previous section, up to nine secondary authors are included in the primary author physical record, and primary author is given in each secondary author record.

COMPENDEX

Authors' names are represented as SMITHbJS in variable length fields, each author name being preceded by a three-digit code with a value of 201 to 299. Thus, up to 100 authors are given.

Bibliographic Citation

Condensates — The Condensates logical record includes the bibliographic information in two physical records. One record contains the short journal title, and the other contains the CODEN (5 characters plus a check character), journal volume (4 characters), journal issue (4 characters), journal year (2 characters), starting page (4 characters) and ending page (4 characters).

POST J — Same as Condensates.

CBAC — Same as Condensates except that there is no check character following the CODEN.

BA Previews — The bibliographic information is contained in two separate physical records.

The first contains:

CODEN (5 characters) followed by a slash

Abbreviated journal title

The second contains:

Accession number (4 characters)

Journal volume number (2 characters) followed by ⌀. Journal issue number in parentheses followed by period and ⌀ (6 characters). If there is no volume, but issue only, then issue information appears in volume place, but still in parentheses.
Journal pages: first page (4 characters), last page (4 characters) and journal year (2 characters). Pagination and year are separated by hexadecimal "FF".

ISI Source

Abbreviated journal title (11 characters)
Journal volume (4 characters)
Journal page, first page (4 characters)
Journal year (2 characters)
Type of source item (one character code) as follows:

- ⌀ - Articles, reports, technical papers, etc.
- A - Abstracts of published items
- B - Book reviews (including critical reviews of books, films, articles, etc.)

C - Corrections, errata, etc.

D - Discussions, conference items

E - Editorial and editorial-like items

I - Items about individuals

L - Letters, communications, etc.

M - Meetings, proceedings from

N - Notes, technical

P - Patents

Q - Bibliography of source items

R - Reviews and bibliographies

COMPENDEX

Abbreviated journal title (variable)

Journal volume (4 characters maximum)

Journal issue (4 characters maximum)

Journal year (4 characters)

Journal pages: first page (4 characters maximum) dash (1 character), last page (4 characters maximum)

Note: The short journal title used by CAS in Condensates, POST J and CBAC is consistent among these data bases, but is not consistent with the abbreviated journal titles used in BA, BioRI, ISI, and COMPENDEX, nor are those three consistent among themselves.

Abstract

Condensates	None
POST J	None
CBAC	None
BA Previews	None
ISI Source	None
COMPENDEX	Full text abstract is contained in a separate field and is preceded by 40%

Title

Condensates	Full titles are give in one physical record and are followed by two equal signs.
POST J	Full titles are given in one physical record and are followed by a period and two equal signs.
CBAC	Full titles are given but they may be broken into more than one record if they exceed maximum record size or if a period occurs as a sentence delimiter within the title. Where this occurs the period is followed by an equal sign.
BA Previews	Titles are enriched, i.e., informative phrases and terms are included in and/or added to the title per se.

The entire enriched title is given in one physical record and is followed by a slash. The end of the title proper is separated from the beginning of the augmenting words by hexadecimal FF.

ISI Source

Titles are given in the last field in the physcial record, which has a maximum length of 170 characters. It includes secondary authors (up to 9) followed by the title. Secondary authors use from 0 to 99 of the 170 characters and the rest are used for the title. If the title exceeds the maximum, the remainder is contained in one or more trailers. If the title does not reach the 170 character maximum length, the record is shortened to the nearest byte divisible by six.

COMPENDEX

Titles are given in a variable length title field. They are included in their entirety. The title field is immediately preceded by 00½.

Research Site

Condensates	Research site of primary author, corporate division, city and state are given.
POST J	Same as Condensates, contained in parentheses and followed by period. If information is not available the code NA is used.
CBAC	Same as Condensates
BA Previews	None
ISI Source	None
COMPENDEX	None

CAS Registry Number

CAS Registry Numbers are included only on POST J and CBAC. The Registry Number is a 9-digit number and provides unique identification of chemical compounds.

Molecular Formula

Molecular formulas are included only on POST J and CBAC tapes.

Compound Name

POST J frequently provides a preferred compound name in a separate physical record. This name may also appear as a keyword in the keyword field. Registry number, molecular formula, and compound name are included for each registered

compound in a given citation.

CROSS Index and Biosystematic Index

CROSS and Biosystematic Index codes are 5-digit codes specific to BA Previews. Multiple CROSS Index codes are separated by blanks and/or asterisks and/or hyphens. An asterisk indicates that the preceding code is a major assignment, while a hyphen indicates that the preceding code is a secondary assignment. Multiple Biosystematic Index codes are separated by blanks and/or asterisks. An asterisk indicates that the preceding code is a new taxa.

Keyword, Index Term, Sentence or Phrase

Condensates	Separate terms and/or phrases.
POST J	Entire text broken up into phrases and written in separate records. The last phrase of each text sentence is followed by.==
CBAC	Entire text broken up into phrases and written in separate records. The last phrase of each text sentence is followed by .==
BA Previews	None (see TITLE)
ISI Source	Subject headings are included in two separate fields.

COMPENDEX — The first occurrence is after the title field and is preceded by 09¢. The second occurrence is after the abstract field and is preceded by 60¢. Keywords referred to by COMPENDEX as "access" words are included in a separate field and are in the last field in the record. Each access word is preceded immediately by a 3-digit number running from 650 to 6[illegible]

CONCLUSION

The advent of document data bases in machine-readable form is a welcome development in information science. These data bases provide an additional information vehicle, one which can be manipulated by fast and relatively inexpensive computer techniques. Although, in many cases, they were developed for such purposes as photo-composition and only come to the information scientist incidentally, they have already proved to be of value in information storage and retrieval of the vast and ever-increasing world of scientific literature.

The conclusion that is obvious from the body of this paper is that the data bases should, from the information

scientist's point of view at least, have been designed specifically for information manipulation, and according to some set of conventions common to all suppliers of the data bases. This was not the case - the data bases were created for purposes unique to each supplier. No set of conventions exists, even in the world of publications from which the machine-readable data bases evolved.

The need for a set of conventions, standards, if you will, is more immediate for machine manipulation of information. When he reads an abstracting journal, the scientist is not greatly bothered by the order of title and author, or whether the author's name is presented as John Jones or Jones, John. He may prefer one way or another, but the variation from journal to journal does not cause him any real inconvenience, only, perhaps, a bit of irritation. When, however, the scientist wishes to use a computer program to scan the abstracts for his interests, these "minor" variations cause very real, very expensive problems. To search two data bases requires two sets of programs, and programs are not generated inexpensively. Those organizations, such as IITRI, that have established search capabilities, mitigate the problem by using techniques such as pre- and post-processors to minimize the extent of program variation. However, as the number of such service centers increases, there will be a great deal of duplicative effort in reformatting machine-readable data bases in order to provide

information services from them. This effort would be better done by the suppliers in preparing their data bases for distribution according to a common set of conventions.

The data bases vary. The requirements for internal processing by the suppliers vary. But a reasonable goal would be the definition of a distribution standard to which all suppliers would convert their data bases. The guidelines currently embodied in the MARC II format of the Library of Congress (the MARC--Machine Readable Cataloging -- format conforms to the as yet unofficial USASI Standard for bibliographic citations and thus is similar to the COSATI format and others), would serve as an initial goal for the suppliers. This extremely flexible format could probably serve them all. Since it is so flexible, all problems of data base variation would not be solved, but a big step would have been taken. At this point the suppliers could get together regarding such items as how to represent an author's name, what type of data comprise a bibliographic citation, etc. These problems are knottier than physical tape format, but any agreements reached would be a boon to the information scientists. Cooperation in both of these areas is a necessity for allowing machine-readable data bases to reach their fullest potential as tools for science. The Standards Committee of ASIDIC (Association of Scientific Information Dissemination Centers) is currently working with tape supplier and center representatives in an effort to ease some of these problems.