IDRC
CRDI
CIID

CANADA

NEW HORIZONS IN AGRICULTURAL INFORMATION MANAGEMENT

PROCEEDINGS

OF AN INTERNATIONAL SYMPOSIUM

MARCH 13-16, 1991

BEIJING, CHINA

The International Development Research Centre is a public corporation created by the Parliament of Canada in 1970 to support research designed to adapt science and technology to the needs of developing countries. The Centre's activity is concentrated in six sectors: agriculture, food and nutrition sciences; health sciences; information sciences; social sciences; earth and engineering sciences; and communications. IDRC is financed solely by the Parliament of Canada; its policies, however, are set by an international Board of Governors. The Centre's headquarters are in Ottawa, Canada. Regional offices are located in Africa, Asia, Latin America, and the Middle East.

Le Centre de recherches pour le développement international, société publique créée en 1970 par une loi du Parlement canadien, a pour mission d'appuyer des recherches visant à adapter la science et la technologie aux besoins des pays en développement; il concentre son activité dans six secteurs : agriculture, alimentation et nutrition; information; santé; sciences sociales; sciences de la terre et du génie et communications. Le CRDI est financé entièrement par le Parlement canadien, mais c'est un Conseil des gouverneurs international qui en détermine l'orientation et les politiques. Établi à Ottawa (Canada), il a des bureaux régionaux en Afrique, en Asie, en Amérique latine et au Moyen-Orient.

El Centro Internacional de Investigaciones para el Desarrollo es una corporación pública creada en 1970 por el Parlamento de Canadá con el objeto de apoyar la investigación destinada a adaptar la ciencia y la tecnología a las necesidades de los países en desarrollo. Su actividad se concentra en seis sectores: ciencias agrícolas, alimentos y nutrición; ciencias de la salud; ciencias de la información; ciencias sociales; ciencias de la tierra e ingeniería; y comunicaciones. El Centro es financiado exclusivamente por el Parlamento de Canadá; sin embargo, sus políticas son trazadas por un Consejo de Gobernadores de carácter internacional. La sede del Centro está en Ottawa, Canadá, y sus oficinas regionales en América Latina, Africa, Asia y el Medio Oriente.

This series includes meeting documents, internal reports, and preliminary technical documents that may later form the basis of a formal publication. A Manuscript Report is given a small distribution to a highly specialized audience.

La présente série est réservée aux documents issus de colloques, aux rapports internes et aux documents techniques susceptibles d'être publiés plus tard dans une série de publications plus soignées. D'un tirage restreint, le rapport manuscrit est destiné à un public très spécialisé.

Esta serie incluye ponencias de reuniones, informes internos y documentos técnicos que pueden posteriormente conformar la base de una publicación formal. El informe recibe distribución limitada entre una audiencia altamente especializada.

IDRC-MR293e
May 1991

New Horizons in Agricultural Information Management

Proceedings of an International Symposium,

March 13-16, 1991, Beijing, China

Compiled and Edited by

Gary K. McCone

IDRC
CANADA

CHINESE ACADEMY OF AGRICULTURAL SCIENCES
SDIC

Sponsored by

International Development Research Centre

Organized by

Scientech Documentation and Information Centre

Chinese Academy of Agricultural Sciences

Organizing Committee

Main Organizing Committee

Chairman	WANG Xianfu
Vice Chairman	JIAO Bin
	HE Chunpei
Secretary	MIAO Zhuoran
Members	HAN Ling
	JIA Shangang
	HUANG Xuegao
	GUO Dianrui
	ZHAO Huaying
	PAN Shuchun
	LI Kaiyang

Secretariat

Chairman	MIAO Zhuoran
Members	CHEN Junying
	HU Jia
	YU Fenghui
	TAI Weidong
	FANG Baoqin

Accomodation & Transport

Chairman	HUANG Xuegao
Members	ZHAO Huaying
	QIN Juanjuan

Conference Site

Chairman	HE Chunpei
Members	PAN Shuchun
	ZHANG Rongchang

Papers

Chairman	JIA Shangang
Members	LI Kaiyang
	LIANG Suzhen
	WANG Zhenjiang
	GUO Jian

Table of Contents

Keynote Address

Session I: Management and Development of National Agro-Information Systems

Session II: Information Management and New Technology Application

Session III: Management and Development of Regional Agro-Information Systems

Session IV: Scientech Information and Productivity

Session V: Development and Utilization of Agro-Information Resources

Appendix 1: Supporting Papers

Appendix 2: Symposium Participants

Appendix 3: Author Index

Database Design at ICRISAT and the Experience of Using External Databases

L. J. HARAVU*

ICRISAT
Patancheru
Andhra Pradesh, India

1. Introduction

This paper outlines what we believe to be general principles of database design, and how these have been applied at the International Crops Research Institute for the Semi Arid Tropics (ICRISAT) for the design of an inhouse bibliographic database, and the experience of using data from two external databases in the creation and maintenance of the inhouse database.

2. Database design principles

2.1 Hardware and software independence

A bibliographic or other database must be viewed as a collection of data elements and relationships between data elements describing real world entities. In the case of bibliographic databases, the entities are books, journal articles, reports, non-book materials, subject terms, names, etc.

The database is not an end in itself but a means to an end. The needs that a database should serve may be said to be:

- it should describe entities (e.g., documents, projects, personnel, etc.) sufficiently adequately,
- the information retrieved from the database should be useful in research, development, problem-solving and decision-making,
- it should provide data for the management and control of functions, and
- it should enable the production of required reports, information products, and services.

Given the above view of a database, it follows that the conceptual design of a database can and should be independent of hardware or software considerations. The database

* Now Head, Information Services, International Livestock Center for Africa (ILCA), Addis Ababa.

needs first to be conceptualized before it can be implemented in a specific hardware-software environment.

Conceptualization of the database includes providing answers to the following questions:

- What types of entities will be described in the database?
- What characteristics of the entities being described are important? Can the characteristics be defined unambiguously?
- How will the characteristics be determined and described?
- What information and processing needs will the database need to satisfy?
- Who will use the database: information specialists, end-users, or both?
- What data-elements need to be shared between processes/functions?

2.2 The database as a central resource

It is essential to think of a database as a single central resource, around which the different information products and services of an information system should be built. The database must contain all the ingredients that go into meeting probable information needs and in producing all routine information products or reports.

The danger of not viewing the database as a single central resource is the need to create more than one file or database to satisfy different needs or applications. A common example is the tendency to treat the library catalog as different from an information retrieval system describing material such as journal articles, conference papers, reports, etc. The result is two databases, each probably following different rules, styles, and standards. From the point of view of a user, he/she will need to use two databases and probably learn two query languages.

Although conceptually a database is talked of as a single resource, in practice it comprises several logically related files. Software systems ensure that so far as the user is concerned the database appears to be a single entity. This is because the data elements constituting the records of the database can further be categorized into files with well-defined relationships between the files. The software system ensures the interaction between files, e.g., when a change takes place in a given file, the software takes care of making the necessary changes in one or more related files. A good software package or application system ensures that the interaction between the different files of a database is completely transparent to the user.

Database design with certain categories of software also involves the definition of the files constituting the database and their interrelationships. By file relationships we mean the extent to which they share data elements and/or exchange data among themselves. For instance, an acquisition system may export bibliographic data into the cataloging system files and vice-versa.

2.3 Data integration vs. functional integration

It is necessary to distinguish between integration of data in a database and integration of functions in a library or information system. The identification of all useful data elements describing different entities, the analysis of their inter-relationships, and the grouping of these into logically related files is data integration. The extent to which data integration is done may vary from system to system. At one extreme, records are created, one or more for each application, with redundant storage of data across files. On the other hand, the designer may have subjected the collection of useful data elements to a critical analysis for relationships before grouping them into logically related files.

Integration of functions, on the other hand, means the capability of a system to perform more than one function, usually in a chain of functions, without redundant entry of data, i.e., with sharing of data between functions, and shared access to files created and maintained primarily for the management and control of a given function.

For instance, in an integrated library system, the acquisitions subsystem would have query access to the cataloging subsystem files for duplication checking or for capture of bibliographic data required, for example, to acquire another copy of a book already held by the library. Similarly, the circulation subsystem may have access to the cataloging subsystem files and vice-versa. Some files, e.g., an authority file, may be shared by more than one subsystem.

In addition to the requirements for sharing of data and files between functional subsystems, an integrated library system may also need to draw data from an external database. For instance, a cataloging or acquisitions subsystem should be developed so that it could utilize MARC records available on a vendor's system or on a database on a library network (or cooperative).

There is a close relationship between data integration and functional integration. Traditionally, library functions were automated as separate, stand-alone applications. Attempts were made subsequently to integrate the different applications into a single system. In many cases, this meant substantial re-design of one or more applications since the original design did not take into account the close interrelationships between functions and the possibility of shared data and shared access to files of these functions.

3. Internal and external factors in database design.

In addition to considering the above mentioned broad principles of database design, it is important that the internal and external factors, i.e., the environment in which a database exists, should also be taken into account in database design. The following are considered important:

3.1 Internal factors

- kinds of bibliographic or other entities that are considered important in a given organization,

- data elements that are special to the organization planning the database,
- searchability, processing, and output products and services that are considered necessary and useful in the given environment,
- the volume of input data and the expected yearly growth of the database,
- the skills available for description and subject characterization of input items,
- the hardware and other infrastructural facilities available, and
- the software available and its capabilities, e.g., its database definition, search, and report generation capabilities.

3.2 External factors

- exchangeability, actual or potential, of the database with other organizations, and
- derivability of the database from one or more global or regional databases, i.e., the possibility that the inhouse database may draw information from one or more external databases.

4. Design decisions at ICRISAT

4.1 Hardware and software independence

In its present form the database is envisaged as a single flat file consistent with the software package that was available, viz., BASIS. In the event that this needs to be implemented on a different category of software package (e.g., Relational database management system), the relationships between different data elements would need to be more explicitly defined than required by traditional information retrieval (IR) software packages.

The structure has, however, been implemented using a mainframe software package called BASIS developed by Battelle as well as on Micro CDS/ISIS.

4.2 As a central resource

The database at ICRISAT in its present form is considered to be a central resource having the following characteristics:

- it integrates both conventional library material (i.e., monographs), as well as other documentary units such as journal articles, conference papers, book chapters, etc. Further, there is provision to describe different kinds of bibliographic entities (monographs, journal articles, theses, conferences, patents, standards, reports)
- the database together with a suitable end-user interface (which is still to be developed) is also considered as being a potential online public access catalog (OPAC) apart from its obvious use as an information retrieval system

- the following output products are generated from the database:
    - catalog cards for monographs
    - monthly accession lists
    - SDI outputs
    - on-demand search outputs
    - current-awareness lists
    - ad-hoc bibliographies

In its present form, however, the database is not designed to support functional integration. A conscious decision not to go in for an integrated system was taken because of factors in the internal environment. The overriding reason for the decision not to integrate functions was the fact that at ICRISAT, the mainframe computer (a VAX-11/780) was already overloaded, and an integrated system which would necessarily have to be based on the mainframe because of the size of the database, and the need for customer access to the database, would have resulted in a system with an unacceptably poor response time and an even more overloaded system. Further, disk storage on the mainframe is a premium resource at ICRISAT and not enough of it was available for the development of an integrated system.

We believe that our experience is probably true of most developing countries where the mainframe computing resources required to develop an integrated system are generally not available. In the developed world this is not the case since most libraries talking about integrated systems are those that have computers dedicated to library automation, and further they have the benefit of vendor systems that provide them access to centrally and cooperatively created cataloging data. This further points to the need in the developing world to adopt a different strategy for integration. Microcomputer-based Local Area Networks (LAN), read-write optical disks, and CD-ROM databases offer the possibility for developing countries to think of truly integrated systems. Until such a time, laterally integrated systems, i.e., systems that are functionally disparate but which exchange data through special programming, seems to be one way of achieving a measure of integration.

4.3 Use of external databases

A third important decision taken at ICRISAT was that the inhouse database would capture its information from two global databases, viz., CAB International (CABI) and the International Information System for the Agricultural Sciences and Technology (AGRIS) of the U.N.'s Food and Agriculture Organization (FAO). The rationale for the choice of these two databases is that the former covers the conventionally published literature of agriculture quite comprehensively and has high quality abstracts, while the latter also covers nonconventional literature of agriculture particularly from the developing world due to the fact that it obtains its input from national input centers all over the world.

It was decided that we would integrate data from the external databases with locally generated input into a single database to serve multiple end-uses including that of an Online Public Access Catalog (OPAC) at a future date. The decision was taken only after a careful examination of one of the internal factors mentioned in Section 3.1, viz., the expected volume of input data with which we would need to contend. It was found that on average the relevant new input each month from the two sources would be about 400 records. This volume of input was just about what we could manage with the existing professional staff at the ICRISAT library and Documentation Services Division.

The two external databases have different structures, follow different styles, and use different rules for bibliographic description, and hence integration of data from these two sources requires some human intellectual effort, and this is directly proportional to the volume of input that is required to be added each month.

By integration is meant not only the merging of data from the two sources but also ensuring that the data so merged are internally coherent and mutually consistent. In other words, data in some fields are reformatted to take care of differences in the bibliographic description rules; incomplete data in fields of the external databases are completed to the extent feasible; and missing elements added wherever necessary. More importantly, an effort is also made to augment the indexing of the items drawn from the two databases. The augmentation is more to slant the indexing of items where required, given the better understanding that we have of the clientele that is required to be served.

In actual practice, the AGRIS tape received each month is subjected to a selection operation to identify those records that are of interest to ICRISAT. A computer program scans the AGRIS tape to do this. The subset of AGRIS data is simultaneously converted into a fixed format file called a FORMS file whose structure is defined in the BASIS data definition language (DDL), a prerequisite to the establishment of a database using the BASIS software. Once this is done, the data can then be loaded into the BASIS database. The program not only restructures AGRIS data to conform to the inhouse database structure but it also looks at specific fields/subfields to extract data that should go into other fields of the inhouse database. For instance, information on the affiliation of the first author of an item appears in the author field of the AGRIS record within parentheses. The inhouse database, on the other hand, has a separate field for affiliation. The computer program written to create an AGRIS subset looks for the affiliation field in the AGRIS record which is then transferred to the appropriate tagged field of the inhouse record. A similar operation is performed with the CABI tapes received each month. A different computer program is used to restructure the CABI data.

In writing the computer programs an attempt has been made to minimize the manual editing effort required to transfer data from two sources into the inhouse database. The computer program cannot take care of differences in style that exist in the two databases, and this is addressed manually. Proof copy of the data from AGRIS and

CABI is edited by professional staff and corrections are made to the FORMS file before it is loaded into the inhouse database.

4.4 Inhouse database structure

Since it was decided that the inhouse database would be built with data from the two external databases, the structuring of the inhouse database required an examination of the structures of the two external databases.

In practice, the structure of the inhouse database was designed *a priori* using the Reference Manual for Machine-readable Bibliographic Descriptions (Dueruckx, 1981) as the source format for the identification of mandatory and optional data elements (fields and subfields). The data element definitions as found in the Reference Manual were used as the standard for the inhouse database. Once the initial structure was designed, the worksheets of AGRIS and CABI and their database production manuals (Martinelli, 1979; CAB International, 1986) were examined. The purpose of the examination was to find out the following:

- common data elements
- unique data elements
- differences in bibliographic description between the two databases
- differences, if any, in data element definitions

Based on the above examination, useful data elements not covered by the Reference Manual were added. One example of a field that was added is the variant author field in use in the CABI database. It was felt that this was a useful data element to be added to the inhouse database considering the wide variations that exist in the rendering of names of Indian, Arabic, European, and other regions. The variant author field which is also indexed enables author searches to be more flexible.

Following the addition of fields that were required for other purposes (e.g., the date of data entry field required to partition the database for SDI services, and the index string field which holds a string of index terms to be used in producing articulated indexes for a bibliography or current awareness list), equivalence matrices were drawn up to determine the kind of mapping between the fields of the two external databases and the inhouse database.

The mappings fall into one of the following kinds: one-to-one, one-to-many, many-to-one, many-to-many or one-to-none. The one-to-one mappings are obviously the simplest to handle. The handling of other kinds of mappings depends on the way the source database distinguishes components or subfields of a given field. If the subfields in the source structure are explicitly demarcated, then it is possible for a computer program to identify the fields/subfields and distribute them unambiguously into one or more fields or subfields of the target structure. However, this is not often the case. An example is the way the CABI database treats conference documents. There is no way that a computer program can determine where the conference was held, and the dates of the conference. One-to-many mappings in such cases would require manual editing effort.

The mapping matrices together with information on the differences in bibliographic description helped in writing the computer programs to restructure and re-format data in the source structure to the target structure, i.e., the inhouse database.

4.5 Database production manual

It was considered essential to develop a database production manual for the inhouse database. The manual defines different data elements, provides examples, and rules for bibliographic description of various elements. The manual is used as a reference both when creating local input as well as when editing input from the two external databases. The manual has also helped in creating authority files for names of persons, corporate bodies, and titles of serials. It is envisaged that these authority files will become computer-resident and will be used to automatically validate data received from the two external source databases as well as for data added locally.

4.6 Local input

An important added value to the database comes from locally generated input. This includes bibliographic data for books, reports, and other monographic material added to the library. Nonconventional literature accessed by the library by virtue of its exchange relationships with several libraries in the semi-arid tropics (SAT) is an important source of locally generated input. Formal and semi-formal literature originating within ICRISAT is another item of input added locally. Very importantly, each document added to the library is examined to see if analytics (chapters or papers) from the document should be added to the database. We believe that it is this kind of input that adds real value to an inhouse database, since most such information is not available in the external databases. In terms of volume, locally generated input at ICRISAT accounts for 20% of the total input.

4.7 Subject accessibility

It is well known that the two external databases use different thesauri to index their records. A conscious decision was taken to distinguish terms of the *CAB Thesaurus* from that of the *AGROVOC* of AGRIS in the inhouse database. Two separate fields called descriptors and identifiers have been designated to hold *CAB Thesaurus* and *AGROVOC* terms respectively. Local input is indexed using the *CAB Thesaurus* as the control vocabulary. Here again the lack of disk space has not enabled the mounting of a machine-readable thesaurus that could be used to validate input.

Search strategies and user profiles need to recognize the availability of terms from the two vocabularies. This does indeed put a burden on the information specialist searcher, and would probably be unacceptable in non-delegated, end-user searching. However, data in the descriptor and identifier fields can both be mapped into a single field for the purposes of a search or in a user profile. This kind of mapping is allowed by the software used. Given this facility, the fact that there are terms from two terminologies can be made transparent to the end-user. Conceptually, it is possible to think of a user-interface which will provide access to terms of both thesauri in an integrated manner.

Given the seemingly unalterable fact that more than one global database in the area of agriculture will continue to exist, and the fact that they will continue to use different vocabularies, it becomes necessary for local systems such as that at ICRISAT to contend with the inevitable differences by seeking software solutions that to some extent minimize the disadvantages of separate rules for bibliographic descriptions, and different vocabularies for subject characterization. The database at ICRISAT and the methodology for its creation and maintenance is an attempt in this direction.

4.8 Duplicate checking

It is well known that there is overlap in the coverage of the two external databases (Deselaers, 1986), and that several common records exist in the two databases. The elimination of duplicate records has been done using a feature of the BASIS software which allows the definition of a duplicate check key in the Data Definition. This key is automatically created for all records that are added to the database, and the key for new records is matched against that of the older records. Potential duplicates, if any, are flagged. These are then checked manually before being discarded or retained.

We have not discovered any apparent pattern in the occurrence of duplicate records in the two databases each month. This is understandable considering that we are concerned with a very small subset of each of the two databases. As a rule, when duplicate records do arise, the CABI record is retained and the AGRIS duplicate is discarded. This is because of the abstracts in the CABI records.

4.9 Exchangeability

One of the main reasons for using the Reference Manual as the source format for the inhouse database was to ensure that it has at least a minimum set of data elements for the identification and description of different types of bibliographic records. Also, conforming to standard data element definitions and bibliographic description rules, we believe will enable us to provide our search output or subsets of our database to interested institutions in an internationally accepted exchange format. Computer programs have already been written to convert records in the BASIS database into a ISO 2709 formatted file. It is also possible, given the tagging scheme of an institution, to download our records to conform to the scheme of the institution wishing to receive our records. The idea is to provide subsets of our database to interested agricultural research stations in the semi-arid tropics (SAT), especially in Asia and Africa, for possible use with Micro CDS/ISIS. Although we have not attempted it, we do not think it would be difficult to convert our records into the structure of the Common Communications Format (CCF).

4.10 Some statistics

Table 1 provides some data on the extent of use of the two external databases and internal input in the creation and maintenance of the inhouse database.

Year	AGRIS input	% of total	CABI input	% of total	Local input	% of total	Total
1988	1269	26.45	2322	48.40	1207	25.16	4798
1989	1758	25.32	3822	55.04	1364	19.64	6944
Total	3027	25.78	6144	52.32	2571	2190	11742

Table 1.

5. Conclusion

A methodology for building an inhouse database using subsets of two external databases has been conceived and implemented at ICRISAT. We believe that the methodology is suitable for special libraries and information centers operating in specific and well defined areas of agriculture or in mission-oriented organizations. The methodology is probably not suitable for information systems operating in relatively broad areas of knowledge since the human effort involved in integrating external data into the inhouse database will then become substantial, and hence also the cost of building the inhouse database.

We see the following advantages in the building of databases using the methodology described:

- The inhouse database has the potential to become, in the long run, more comprehensive than either of the one or more databases from which the inhouse database draws its information, in the specific areas of interest of a given information system.
- Value addition to the inhouse database is possible in several ways:
    - through local input, e.g., nonconventional literature,
    - through inclusion of other kinds of information, e.g., on specialists, research projects,
    - through slanted indexing.
- Online access to a comprehensive information resource to end-users within the organization through a well designed user-friendly interface at considerably less cost than would be possible by other means. Quite conceivably, as telecommunications in the developing world improve, such databases can be remotely searched from institutions within the country. At ICRISAT, we hope to provide online access to our database to remote users in India. With the emergence of WORM media, there is also the possibility of distributing such databases on optical disks.

- SDI and search services to users based on the inhouse database would have the advantage of wide coverage from a single source at considerably less cost than possible otherwise.

In most developing countries, especially in Asia and Africa, where the telecommunications infrastructure is still too poor or expensive to think of providing end-user access to remote databases outside the country, it would appear that the building of local databases in specific areas of interest is a way out of the situation, until databases on optical disk cover more areas and become available at considerably less cost than they are today.

We believe that an inhouse database should be seen as a repackaged product that not only brings the global database closer to the user but one that adds value by the inclusion of new information and the capability to utilize the information in friendly and innovative ways.

The existence of an inhouse database does not in any way eliminate the need for access to external databases. However, such access will probably be required in the interface and interdisciplinary areas, e.g., biotechnology, food, nutrition, etc., with the inhouse database becoming the mainstay for information retrieval and dissemination services.

6. References

Dueruckx, H., and Hopkinson, A. *Reference manual for machine-readable bibliographic descriptions*. 2nd revised ed. (PGI-81/WS22). Paris, France: Unesco, 1981.

Martinelli, M.T. *AGRIS: guidelines for bibliographic description and input sheet preparation*. FAO-AGRIS-4 (Rev. 2). Rome, Italy: FAO, AGRIS Coordinating Center, 1979.

CAB International. *CAB abstracts database production manual*. Farnham Royal, Slough, UK: CAB International, 1986.

Deselaers, N. The necessity for closer cooperation among secondary agricultural information services: an analysis of AGRICOLA, AGRIS and CAB. *Quarterly Bulletin of the International Association of Agricultural Librarians and Documentalists* 31(1): 19-26, 1986.