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.

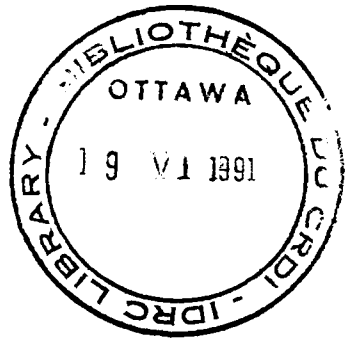
Esta serie incluye ponencias de reuniones, informes internos y documentos técnicos que pueden posteriormente conformar la base de una publicación formal. El informe recibe distribución limitada entre una audiencia altamente especializada.

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

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

On the CAB Thesaurus

HOU Hanqing XU Jia

Department of Library and Information Sciences
Nanjing Agricultural University
Nanjing, China

Abstract

This paper presents an outline of major thesauri in agriculture all over the world, analyzes the characteristics and position of the CAB Thesaurus (CABT) and concludes that it is the thesaurus of the largest scale, and the widest influence as well as having many distinguishing features in the field of agriculture. The paper points out that the thesaurus integrates an alphabetical display with the functionality of a hierarchical index, a permuted index and an index of identifiers and gives a complete display of hierarchical relationships in its alphabetical list, which breaks traditional models of thesauri. Moreover it observes that while this new kind of structure is concise, practical, convenient to use, and is one from which people can obtain a great deal of information by one look-up, the structure makes the thesaurus too lengthy to retrieve from easily on the basis of disciplines and specialties. The paper also evaluates the performance of the CAB Thesaurus with three standards (i.e., connectedness ratio, accessibility measure, and equivalence ratio) and indicates that the structure of the thesaurus is adequate, but non-descriptors in it are insufficient. In displaying hierarchical relationships and preferring the popular name over the Latin scientific name, it obviously violates international standards for the establishment of thesauri, which may cause confusion to users.

CAB ABSTRACTS and the CAB database are the most common information retrieval tools used by agricultural scientific and technical researchers in China. However, the *CAB Thesaurus* (*CABT*), as the indexing language of CAB ABSTRACTS and the CAB database, is less well-known in China, even the agricultural libraries and information services know little about it. This paper intends to introduce the characteristics of *CABT* and to evaluate its structure and cross references so as to provide guidance to numerous agricultural scientific and technical researchers and to those who will themselves establish thesauri.

1. Characteristics and position of CABT

To meet the needs of indexing and retrieving agricultural documents, some agricultural information agencies in European and American countries, international agricultural libraries and information services have established and published a number of thesauri covering agriculture and its related disciplines since the 1960s. According to *Thesaurus Guide: An Analytical Directory of Selected Vocabularies for Information Retrieval* there are eighteen English language thesauri in agriculture and its related disciplines which are available now in the world. Among these, there are five which include over 10,000

terms, 11% of the total number of large English thesauri included in *Thesaurus Guide*. These are as follows:

(1) *A Multilingual Thesaurus of Agricultural Terminology* (*AGROVOC*), compiled by FAO and CEC, 1982, 1st ed., 7 vols, number of descriptors: 8,660, number of non-descriptors: 6,940.

(2) *Agricultural/Biological Vocabulary* (*A/B Vocabulary*), compiled by NAL of USDA, 1967, 1st ed., 2 vols, number of descriptors: 11,800, number of non-descriptors: 6,700.

(3) *CAB Thesaurus* (*CABT*), compiled by CAB, 1983, 1st ed., number of descriptors: 40,800, number of non-descriptors: 7,200. 1988; 2nd ed., number of descriptors: 47,400, number of non-descriptors: 8,600.

(4) *Canadian Agriculture Thesaurus* (*CAT*), compiled by Agriculture Canada, Research Branch, Research Program Service, 1977, 2nd ed., 4 vols, number of descriptors: 24,000, number of non-descriptors: 12,000.

(5) *FAO Index Terms*, compiled by Documentation Processing Section of FAO, 1981, 2nd ed., 2 vols, number of descriptors: 14,000, number of non-descriptors: 1,000.

From what is listed above, we can clearly see that *CABT* has the most important position among *CABT*, *AGROVOC* and *Agriculture/Biology Vocabulary* which are the most famous agricultural thesauri in the world.

CABT is the largest agricultural thesaurus in the world at present. The number of terms in *CABT* is much greater than the sum of *AGROVOC* and *A/B Vocabulary*. The first edition contains 48,000 terms, in the second edition published in 1988, the number has gone up to 56,000. According to statistics from the *Thesaurus Guide*, there are eleven English language thesauri containing more than 20,000 terms, *CABT* stands in the top position, other famous large thesauri such as *TEST*, *Thesaurofacet*, *MeSH*, *INIS*, *EURATOM*, and *NASA*, etc., fall behind it.

CABT has the broadest subject scope among all agricultural thesauri in the world. It covers agriculture and almost all of its related disciplines. Besides agricultural science, forestry, and animal husbandry, it contains agricultural economics, soil technology, aquaculture, veterinary science, food and nutrition, agricultural environmental pollution, agricultural education and extension and so on.

The Commonwealth Agricultural Bureaux (International) is the greatest agricultural information service in the world. It publishes 46 abstract journals every year, in which about 150,000 papers in agriculture collected from many countries in the world are reported. The starting point for the *CAB Thesaurus* was the *CAB ABSTRACTS Word List* (1978), drawn up on the basis of indexing and retrieving over 2,000,000 agricultural articles and the frequency of occurrence of terms used in the subject index of the CAB abstract journals. And then, with reference to FAO's *AGROVOC* and the CEC's

thesauri on veterinary science and agricultural economics, CAB published *CABT* in 1983. Thus it has literary warrant and meets the demands of agricultural information work.

CABT is of wide adaptability and strong practicality, so it is well appreciated by numerous agricultural scientific and technical researchers. It has great influence in English-speaking countries as well as non-English countries. In 1985, NAL of the U.S. began using *CABT* as the indexing language of its AGRICOLA database. An additional 1,500 terms for subjects covered by AGRICOLA but out-of-scope for CABI (mainly human ecology, food service, and food technology) were admitted into *CABT* at this time and each was marked so as to be identified by users. Thus *CABT* has been widely employed by two of the three largest agricultural bibliographic databases in the world. The important position of *CABT* is further affirmed in agricultural information services, and it will have greater and greater influence on the indexing and retrieving of agricultural literature.

2. The structure of CABT

The structure of thesauri has gone through a progression from simple to complex, and has gradually formed a basic model, i.e., a thesaurus consists of an alphabetic list, a category list and a hierarchy index (or hierarchy graph). In addition to all these components, some thesauri contain an index of identifiers, a permuted index, a bilingual concordance or an alphabetic index, etc. The structure of thesauri therefore has become more and more complicated. For users who want to look up a descriptor and the hierarchy in which it belongs, it is necessary to search more than one part of a thesaurus, which makes the speed of indexing and retrieving slow down.

Since the 1970s, the designs for thesauri have tended to a new direction, i.e., the functionality of thesauri has strengthened and the structure has simplified. Besides *The Thesaurofacet* established by Jean Aitchison in 1969, *CABT* is another typical example of this new tendency.

The entire *CABT* has only an alphabetic list without any other components. It displays all hierarchical relationships in the alphabetic list, which actually means that the hierarchy index is integrated into the alphabetic list. It also includes identifiers and arranges them together with common descriptors, so a separate index of proper names can be omitted. Thus it integrates aspects of a hierarchy index, an index of identifiers and a permuted index within the alphabetic list and forms a new kind of structure.

Before *CABT* was published, some thesauri such as *TEST* (1967), *NASA* (1967), and *JICST* (1975) had employed the style which displays all hierarchical relationships, but they did not break the traditional model at all, they retained either a hierarchy index or permuted index and category index. However, *CABT* boldly abandons the hierarchy index, permuted index and category index, and draws the quintessence from various thesauri and then makes some improvements in the style of display. By displaying all hierarchical relationships in its alphabetic list, the hierarchical structure of *CABT* is so

distinct and directly perceived that users can conveniently find all the superordinate and subordinate terms under one term, then choose any of them. The unitary structure consisting of an alphabetic list as mentioned above is convenient to master and use, in one look-up users can obtain a maximum of information and avoid searching in several parts of the thesaurus.

However, the structure and the display still have three problems as follows:

1) **The thesaurus is lengthy and costly.** Although it omits a hierarchy index, a permuted index and some other components, its size is still greater than that of traditional thesauri. Along with the extension of subdivisions of hierarchy and more terms being included in the thesaurus, the size of a thesaurus displaying all hierarchical relationships will be larger and larger.

2) **A whole hierarchy can be displayed under its top term, but the hierarchy can not be displayed completely under any other terms.** So if users want to know a whole hierarchy clearly, they must seek under its top term. For example, under the term "maize," the broader term "cereals" and the narrower term "sweetcorn" can be found, but some other terms which belong to the same hierarchy such as "wheat," "rice," and "sorghum" can not be found. Under the term "sweetcorn" even "flint corn," and "dent corn" can not be found. Only under the top terms such as "cereals" or "forage crops" can the whole hierarchy be displayed.

3) *CABT* **only has an alphabetic list without a classified list.** So it is not convenient to browse various subjects on the basis of disciplines and specialties and to retrieve on a large-scale. All the CAB ABSTRACTS journals have been set up under a broad classification scheme. By gathering and collating all these classification systems, a classified list could probably be established.

3. The cross-reference system of CABT

CABT (2nd ed.) consists of 56,000 entries. The chief component of each entry is its cross-references. Besides those, a few of the entries contain scope notes. The cross-reference system of *CABT* includes USE, UF, BT, NT and RT, by which forms of terms, meanings of terms, and inter-term relationships can be effectively controlled. Besides the above-mentioned five kinds of cross-references, *CABT* (1st ed. & 2nd ed.) also uses OR cross-references for individual cases.

F.W. Lancaster has proposed that a thesaurus can be evaluated by some existing standards, namely connectedness ratio, accessibility measure and equivalence ratio.

The connectedness ratio is the ratio of cross-referenced terms (i.e., terms linked to at least one other term; e.g., by BT, NT, or RT) to total terms in a thesaurus. In other words, the fewer non-cross-referenced terms a thesaurus includes, the higher the connectedness ratio will be, and the better structure the thesaurus will have. According to statistics from a test performed by taking a random sample of *CABT*, there are only

four non-cross-referenced terms in 120 pages of *CABT* (10% of the length of *CABT*, 1st ed.), there are 4,855 terms in these sample pages. So the connectedness ratio of *CABT* is 0.999.

The accessibility measure is the mean number of references received by a descriptor in a thesaurus. The more each term in a thesaurus is referred to other terms, the higher the accessibility measure. This can indicate whether an inter-term relationship in the thesaurus is sufficiently displayed. According to the sampling statistics, the accessibility measure of *CABT* is 5.78, in which the BT-NT accessibility measure is 4.78, and the RT accessibility measure is 1.08.

The equivalence ratio is the ratio of nondescriptors to descriptors in a thesaurus. It relates to the number of nondescriptors in the thesaurus. In the first edition of *CABT*, the number of nondescriptors is 7,200, and the number of descriptors is 40,800, so the equivalence ratio of the first edition is about 0.18. In the second edition, the number of nondescriptors is 8,600, and the number of descriptors is 47,400, as a result the equivalence ratio of the second edition is also 0.18, the same as the first edition. A comparison between *CABT* and several other authoritative thesauri in the world is shown in Table 1.

Name of Thesaurus	Connectedness Ratio	Accessibility Measure	Equivalence Ratio
MeSH (1976 ed.)	0.573	0.381	0.5*
LCSH (7th ed.)	0.799	1.032	0.22*
ASTIA (2nd ed.)	0.956	1.492	-
TEST (1st ed.)	-	2.903	0.31*
CABT (1st ed.)	0.999	5.860	0.18

Table 1. The comparison between *CABT* and several other authoritative thesauri

(The data marked with an asterisk are obtained from the statistics of *Thesaurus Guide*, the data for *CABT* (1st ed.) are based on the sampling statistics and the rest are cited from *Vocabulary Control for Information Retrieval* (Lancaster, 1972).

From Table 1, it can be seen that the connectedness ratio and accessibility measure of *CABT* are both the highest among all the thesauri listed above, which indicates that *CABT* is superior to the other thesauri in the display of inter-term relationships and is therefore of better adaptability. As for the accessibility measure, it is generally considered that the ideal value should range from 2 to 5. As the result of displaying all the hierarchical relationships in *CABT*, the BT-NT accessibility measure amounts to 4.78 and the RT accessibility measure reaches only 1.08. However, the accessibility measure is adequate in general. The equivalence ratio of *CABT* is far less than those of the several other thesauri mentioned above and less than the average of large thesauri (i.e., 0.62) in the 1980s as well. This shows that the number of nondescriptors included in

CABT is not sufficient, which probably relates to the fact that there are few USE cross-references in the subject index of CAB ABSTRACTS. In view of the above-mentioned fact, some synonyms and quasisynonyms, which appear frequently in documents and/or are possibly used by users, should be added to *CABT*. In addition, some excessive specific terms can be replaced by their superordinate terms and then set up USE cross-references between these subordinate terms and superordinate terms. In a word, more entry terms should be added to *CABT* to facilitate indexing and retrieving.

In the areas of collection and selection of terms, vocabulary control, symbols of cross-references, display of terms and their relationships, *CABT* basically follows ISO 2788 (i.e., *Documentation-Guidelines for the Establishment and Development of Monolingual Thesauri*), there are obviously two problems as follows:

1) Treatment of the hierarchical relationships. ISO 2788-1986 stipulates that only those terms with the generic relationship and whole-part relationship which covers a limited range of situations can be organized as a hierarchy.

The structure of *CABT* does not strictly follow this rule. Almost all the terms with a hierarchical whole-part relationship are organized as a hierarchy; even those terms with the whole-aspect relationship or with an associative relationship are related hierarchically. Here are some typical examples:

Blood
BT1 body fluids
BT2 animal anatomy
BT2 fluids
NT1 blood cells
NT2 erythrocytes
... ...
NT1 blood composition
NT2 blood chemistry
NT3 blood lipids
NT3 blood sugars
NT2 blood proteins
RT blood proteins
RT blood sugars

According to ISO 2788, in a whole-part relationship, it is only 1. body and organs, 2. geographical location, 3. administration and social community, and 4. branches of disciplines that we can use to establish BT/NT cross-references as hierarchical relationships. But it is apparently contradictory to regard the descriptors "blood cells," "blood proteins" and "blood sugars" which have a whole-part relation with the term "blood" as related descriptors (RT) and as narrower descriptors (NT) of "blood" at the same time. Facing this kind of processing method, which you can find everywhere in *CABT*, we wonder why the editors regard the same descriptor as NT as well as RT. In addition, it is also not appropriate to set up BT/NT references for non-hierarchical terms "animal

anatomy" and "blood," "blood" and "blood composition," "blood composition" and "blood chemistry." So, the establishment of such disordered relationships, which you can find throughout the thesaurus makes users feel that it is much more difficult to access.

2) In distinguishing popular names and scientific names, ISO 2788-1986 specifies that if a popular and a scientific name refer to same concept, the form most likely to be sought by the users of the index should be chosen. For example "penguins" might be chosen as the preferred term in a general index, but the scientific equivalent, "Sphenisciforms," may be preferred in a zoological index. Reciprocal entries should be made in these cases. The standard also requires that USE/UF references should be established for the two terms. But, for *CABT* it's difficult to solve the problem, for instance, "Field Crops Abstract" often uses popular names to represent "crops," whereas "Plant Breeding Abstract" usually chooses scientific names. In order to avoid disorders, *CABT* provides double-nomination for some kinds of plants which use both popular and scientific names as descriptors, while a RT reference is set up between them. For example, following are the two entries "Zea Mays" and "Maize" for the concept "Maize" in *CABT* (1st ed.).

Zea Mays	Maize
uf corn*	uf corn*
bt1 cereals*	bt1 cereals*
bt1 crop plants as weed	bt1 silage plants
bt1 fodder crops*	bt2 fodder crops*
bt1 oil plants	nt1 dent maize
bt1 Zea	nt1 flint maize
bt2 gramineae	nt1 soft maize
rt alcoholic beverage*	nt1 sweetcorn
rt breakfast cereals*	... ...
rt flint maize	rt alcoholic beverage*
rt maize*	rt breakfast cereals*
rt maize chlorotic dwarf virus	rt cornflour
rt maize dwarf mosaic virus	rt maize*
... ...	rt maize oil
rt popcorn*	rt maize starch
rt starch	rt popcorn*
rt sweetcorn	... ...
rt tassels	rt Zea Mays

In these two entries, "Zea Mays" and "Maize" contain 23 and 47 cross-references respectively. Among them seven of the uf, bt and rt entries are the same, the others are quite different, even contradictive. In addition, the popular name contains some nts while there are not any under the scientific name at all. Some nts such as "flint maize," and "sweet corn" are even put in the rt reference area. Sometimes the referred word's form is different when two entries for descriptors are directed to the same concept when

setting up references. For example, when referred to "starch," one entry is directed to "starch," the another to "Maize Starch." You can find similar results in the survey of four main crops expressed by eight terms in Table 2, which was examined by the authors.

	UF		BT			NT			RT			
	Tot	Sam	Tot	Sam	Dif	Tot	Sam	Dif	Tot	Sam	Sim	Dif
Maize	1	1	3	2	1	8	0	8	35	4	1	30
Zea Mays	1	1	6	2	4	0	0	0	16	4	1	11
Rice	1	1	1	1	0	5	0	5	11	1	2	8
Oryza Sativa	1	1	5	1	4	0	0	0	16	1	2	13
Wheat	0	0	1	1	0	4	0	4	16	0	1	15
Triticum Aestivum	0	0	4	1	3	0	0	0	8	0	1	7
Cotton	0	0	5	1	4	0	0	0	16	0	0	16
Gossypium Hirsutum	0	0	4	1	3	0	0	0	1	0	0	0

	Sum		
	Total	Same	Different
Maize	47	7	40
Zea Mays	23	7	16
Rice	18	3	15
Oryza Sativa	22	3	19
Wheat	21	1	20
Triticum Aestivum	12	1	11
Cotton	21	1	20
Gossypium Hirsutum	5	1	4

Table 2. Comparison of entries of scientific name with common name in *CABT*.

From Table 2, we can see that the process of making separate entries for descriptors under scientific name and popular name of plants and animals artificially cuts apart the related linkage and destroys the integration and consistency in cross references. As a result, some of the cross references are repetitive, and some are not enough, nts are put together under the popular names, whereas rts are collected under the scientific names. It is suggested that the method of selecting "the form most likely to be sought by users" (usually popular names), collecting all references under the popular names and making USE/uf links between the popular names and the scientific names should be used. This method can make references under the descriptors a complete, unified organization, avoid differences from beginning to end, reduce errors and make it easy for users to select what they want to look up. It would also make the abstracts journals which must choose scientific names easier to use.

Although there are some defects, *CABT* is still a thesaurus of great impact, with the most users and the most influence in the world of agriculture. It is hoped that after further improvements, the next edition of *CABT* will become the most commonly used indexing language in agricultural information work.

References

1. GID: Thesaurus guide; analytical directory of selected vocabularies for information retrieval, GID, 1985.

2. Hou Hanqing and Xu Jia : "Survey and trends of thesaurus construction abroad." Qing Bao Xue Bao (*Journal of China Society of Scientific & Technical Information*) Vol. 8, No. 5, 1989.

3. Lancaster, F.W.: *Vocabulary control for information retrieval*, second edition. Washington, Information Resource Press, 1986.

4. *ISO 2788-1986(E) Documentation--Guidelines for the establishment and development of monolingual thesaurus*, second edition, 1986.