IAEA FAO UNESCO UNIDO

INTERNATIONAL SYMPOSIUM ON THE FUTURE OF SCIENTIFIC, TECHNOLOGICAL AND INDUSTRIAL INFORMATION SERVICES

Leningrad, USSR, 28–31 May 1990

IAEA-SM-317/ 21

COMPUTERIZED INTEGRATED DATA BASE PRODUCTION SYSTEM (COMPINDAS)

D. MAREK; DR. K. BÜRK

FACHINFORMATIONSZENTRUM KARLSRUHE
Gesellschaft für wissenschaftlich-technische Information mbH

D-7514 Eggenstein-Leopoldshafen 2
Federal Republic of Germany

IAEA-SM-317/21

COMPUTERIZED INTEGRATED DATA BASE PRODUCTION SYSTEM (COMPINDAS)

D. MAREK; Dr. K. BÜRK

FACHINFORMATIONSZENTRUM KARLSRUHE, Gesellschaft für wissenschaftlich-technische Information mbH, Eggenstein-Leopoldshafen, Federal Republic of Germany

ABSTRACT

Based on many years of experience, and with the main objective in mind to guarantee long-term database quality and efficiency of input processes, Fachinformationszentrum Karlsruhe is developing an integrated interactive data management system for bibliographic and factual databases.
Its concept includes the following range of applications:

- Subject analysis with computer-assisted classification, indexing and translation.
- Technical procedures with online acquisition and management of literature and factual data, recording by means of optical scanning, computer-assisted bibliographic description, control and update procedures.
- Support of the whole process by continuous surveillance of document flow.

All these procedures will be performed in an integrated manner. The system is to meet high standards for flexibility, data integrity and effectiveness of system functions. Independent of the type of data, the appropriate database or the subject field to be handled, all data will be stored in one large pool.
One main goal is to avoid duplication of work and redundancy of data storage.
The system will work online, interactive and conversational. COMPINDAS is being established on the basis of the ADABAS as database management system for storage and retrieval. The applications are being generated by means of aDis of ASTEC in Munich. aDis is used for the definition of the data structures, checking routines, coupling processes, and the design of dialogue and batch routines including masks.

Preconditions, Volume of Activities and General Objectives

FIZ Karlsruhe offers a broad range of services. One of its main activities is the production of databases for all aspects of energy, physics, mathematics and related fields of science and technology. Every year more than 230,000 documents - articles, research reports, books, patents and conference papers - are systematically collected, analyzed and processed. Besides these bibliographic entries, databases comprising about 45,000 entries with factual information are being built and maintained. They include information on conferences, research in progress, institutions, experts or products. The so-called multi-dimensional databases are covering several aspects.

The library catalogue of FIZ consists of 600,000 monographic entries plus 7,000 updates per year and 7,000 periodicals and serials. Data are originating from very different sources and processed under various conditions, e.g. as international cooperative ventures like INIS. From these data about 50 different products like magnetic tape services, reference journals etc. are being created. Due to the requirement of complete coverage for each database, there may exist some intentional overlaps among the different databases.

INIS specifications and principles serve as a model for every database we create. This means that our input procedures follow as strictly as possible INIS standards and rules. We maintain and update computerized thesauri, subject classification schemes as well as authorities for corporate entries, journal titles, country and language abbreviations, etc.

Based on many years of experience, Fachinformationszentrum is developing a new input system for the production of its bibliographic and factual databases. Precondition for the new system was, that it should, as far as possible fit into the existing soft- and hardware surrounding. It should be operated on IBM 3081-KXS mainframe and ADABAS was to serve as a database management system. The changeover should be performed in several steps without interrupting or disturbing production schedules.

Functional Requirements

The most important requirement was that COMPINDAS could be realized as an open easy-adaptable system offering the users a maximum of flexibility and allowing them to cope with the increasing amount of data and any changes in FIZ input tasks and services.

The range of functional activities to be included in the concept are:

- Subject analysis with computer-assisted classification, indexing and translation, as well as interactive maintenance of thesauri and classification schemes.
- Technical procedures with online ordering, acquisition and management of literature and factual data, recording by means of optical scanning, computer-assisted bibliographic description, control and update procedures, as well as creation of different and variable output products.
- Support of the whole process by continuous surveillance of document flow by means of operational statistics as well as handling and accounting service input, including maintenance, adjustment and further development of computer programs.

All the procedures are to be performed in an integrated manner. Operations and methods as well as user's guidance should be compatible. Information must be presented homogeneously. The system is to work online, interactive and conversational. Measures and techniques to guarantee long-term database quality and efficiency of input processes are to be supported.

Concept of the System and File Stucture

According to the concept the new input systems may be regarded as an entity, the pieces of which are the functional activities and user's profiles.
Independent from the subject fields, the kind of data (bibliographic or factual), and the status of processing or the products to be created, all data are being stored in a large complex of files. To achieve an optimum of efficiency duplication of work and redundancy of data storage are being avoided. This of course requires an homogeneous definition of

data elements, a clear file structure and the technical possibility of interconnecting data files and file segments on different levels.

All activities, starting with acquisition of literature and collection of factual data and ending with the creation of the products, will be closely interwoven. Thus e.g. one entry in the documentation file will show up like comprising all information fields needed. In reality they are constructed by links passing over several files. The information elements are recorded when they first appear to be needed. This means e.g., a proceedings volume is described when it is being ordered. The information on the conference itself at the moment may already be present, these data being part of the conference calender file. The same may apply to standardized corporate entries information within the institutions file. When describing one specific article of the proceedings for a database all the information elements already present in the whole system, are linked to form this entry. Only the specific supplementary information needed must be added.

The system will be realized in a totally interactive way. The data recorded are immediately presented for search purposes or to be linked one to another. How is the link perfomed, e.g. to a corporate entry? During the input procedure, while typing the bibliographic description instead of the corporate's name the search terms may be given. When the screen is filled in, and the information sent to the mainframe, not only checking routines but also searches are being performed. If there is one positive and unambiguous answer the data are automatically linked. If there are several possible answers the user is asked to make his choice and to mark the relevant entry. If there is no answer at all he may specify his query or add the new information to the corporate entry file. This new entry, however, will be marked by status 'not yet checked'. The staff responsible for the management of this file may change the status into 'ok' after control. If however there are any remarks to be made to the colleagues this can be done by a kind of simplified mail activity.
Information on the processing status of an entry is - as

already mentioned - part of its description. This allows us to give specific directions for each unique entry.
Consequently, the users will be equipped with laser printers and will be able to create and design printouts and products on their own by means of a user-friendly report generator. This means that they will be free to arrange the process according to their necessities.

Quality control will be suuported by a large range of checking routines to exclude errors during the data collection process already. These are checks on consistency, reasonableness, reliability, logical syntax, duplication and correct spelling.

An important instrument for the preservation of timeliness and completeness are continuous observation of document flow and compilation of statistical data. During the whole process information will be gathered on when the document came in, who is dealing with it just now and so on. This is done by means of a light pen at each terminal to scan the key of the document represented as bar code. The first time the document is in hands it receives its barcode label. This code is also used to call for an entry on the terminal.
The data extracted are processing information as well as information on the kind of the analysis and the products. The users may activate different standard statistical options while performing a range of searches and ask for the data being displayed in several forms including graphs.

Within the system special characters and mathematical formula will be recorded in a linearized form, but they will be shown in a graphics representation to make control and proofreading easier by means of TeX of AMS.

System Components

To realize the system as planned, we first looked for standard systems on the market. The result was that there didn't really exist standard software solving all our tasks, or it was supporting, or documentation, or library tasks only. To adapt such partial systems to your necessities is usually a more labor-extensive task, than to make a completely new one. So we

decided to realize the system on our own on the basis of ADABAS as database management system for storage and retrieval. The applications are being generated by means of the adaptable Documentation and Information System (aDIS) of aStec in Munich. aDIS is used for the definition of the data structures, checking routines, coupling processes, and the design of dialogue and batch routines including masks.
Besides we are using the Statistical Analysis System of the SAS Institute for analysis, evaluation and representation of statistical data.
Generation of the different products and printouts is done by means of the text-formatting-system Con-form from Software AG.

Other components will be already existing autonomous systems supporting data collection, translation and analysis.
By means of an optoelectronic scanning equipment abstracts especially from core journals, will be digitalized. The market offers a wide range of such systems with very big differences in their capabilities. We tested some of them and decided on Kurzweil Discover 7320 which may be operated on IBM PC/AT. One major reason for the decision was its 'Intelligent Character Recognition', which means that Discover analyses the characteristics of a symbol independent from its kind of type. The other reason was the interactive verifier mode, performing a reasonableness check against a dictionary, which is continuously updated from the system by training itself.
For the machine-supported translation of abstracts from German to English we looked at several machine translation systems. Our favorite is METAL (Machine Evaluation and Translation of Natural Language) of Siemens AG. Its basis is an expert system which controls analysis of the whole sentence grammatically by means of rules and a dictionary. The possible interpretations are gathered, compared and weighted following different criteria. The most probable solution is then chosen. METAL runs on SINIX-system with a specific workstation for the programmming language LISP and MX300/MX500 terminals. Our translators are convinced that they can reduce the translation time with the assistance of METAL to one third.
Since October 1985 FIZ KA uses the automated indexing system AIR, which was developed by TH Darmstadt for the machine-supported

indexing of its database PHYS. 120,000 entries a year are indexed by machine in English language. The basis of the system is the hierarchically structured PHYS thesaurus. By analysis of title and abstract text and by means of a knowledge base - a large set of rules and a dictionary - indexing procedure performs a coordinate indexing. The dictionary originally was derived from a set of manually indexed documents and contains weighted information on relation between terms and descriptors continuously updated.

The range of aDIS functions comprises a computer-assisted indexing method as well which, however, is less sophisticated. It is performed by means of the aDIS-subsystem for textual analysis TALSYS for textual analysis. Using a hierarchical thesaurus, which contains supplementary textual information, TALSYS identifies terms to isolate them from their context and interpretes compounds by means of a combinatorial selection routine. In the next step it systematically arranges the words identified and by statistical analysis structures them according to the thesaurus.
Whereas in AIR weighting of a term is done analyzing the complete abstract by copying the human intelligence, TALSYS weights the terms by means of the thesaurus and its hierarchical position only to identify the correct wording.
AIR - with the aim of being more precise - requires the analysis of a human-indexed portion of a database for its dictionary. TALSYS can be used already when a dictionary does not exist. That is the reason why we are planning to use both methods complimentarily. By the same kind of techniques automatic classifying will be realized.

A major problem we had to cope with when planning our system, was word processing, which in a central system will always be a crucial point. We found out that using IBM 3270 terminals we would always be restricted in one or the other way, especially in updating. We thus decided to equip COMPINDAS users with local intelligence, i.e. PC's, and could herewith also solve the problem of showing the users formula not only in a linearized but also in a graphical form.
We are aiming at an integration of all these subsystems in a

most user-friendly way.

Specific Applications and Functions

Realization plan for COMPINDAS comprises 17 packages which will be performed step-by-step following the capacity we have. The first package - the thesaurus maintenance system has already been realized and is offered to the users in a pilot version. The maintenance system comprises all thesauri we use for the different fields including INIS thesaurus, and packs them together by means of a central word list. Each thesaurus with its hierarchical structure is represented individually. The word list is a combination of the words in all thesauri, given in sequential order and supplemented by cardinal form, synonyms, single terms out of components, translations of the terms etc., which may support search, checking and subject analysis. Terms in the wordlist are linked with the appropriate positions within a thesaurus. Searches can be performed via the wordlist or directly in each thesaurus. Updates may be done online or as batch procedures. The thesaurus maintenance system allows you to extract specific thesauri or listings and to establish multilingual dictionaries.

Programming for periodical control subsystem package is just in process. It will allow online ordering of journals, check-in of issues, routing and claim. About 70 - 90 % of our literature input resulting from periodicals, the concept had to concentrate on processes associated with documentation activities. Information on relevance and subject analysis are recorded and used for control or to effect the sequence flow.
Checking of issues is machine-supported by means of a prediction pattern created by statistical analysis of receipt history. The number of relevant articles within an issue will be recorded and leads to the creation of analytical entries linked to the appropriate issue of a journal. Bibliographic analysis of the article will be restricted to the article-specific information.

A major subsystem being just programmed is the management of factual data on institutions and projects. Because of its hierarchical relations and the diversity of products and procedures it is used for, it represents the most complex part of COMPINDAS. It will comprise the hierarchical file showing the organisational structure of the institutions and supplemented with descriptive information, and sequentially arranged files for addresses, country codes and names, persons and projects. All of them may be linked to each other. To create a specific product, components of the whole system are extracted and grouped together according to the specific requirements. The description of an institution may contain as already mentioned the standardized name as represented in the INIS authority file and will be used for checking purposes within the literature file. It may be used as well to create a multidimensional database composed by information on institutions, their research in progress activities and the appropriate publications. In this case the literature file entry is linked to the relevant research-in-progress entry, too.

In parallel to the realization of the different applications a user's manual is prepared which will describe the COMPINDAS concept, the user interface, dialogue and batch functions and the different subsystems.

Future Directions

We are quite aware that realization of COMPINDAS in the form described is only the first step towards an intelligent documentation system which is our final goal. So we will work in parallel to come to an online and conversational indexing and classification system. We also are participating in a research project on computer-supported bibliographic analysis, AUTOCAT. The actual status of AUTOCAT is that optical scanned and digitalized information, e.g. from a journal article is analyzed, subdivided and composed to a bibliographic description. Identification and allocation of the elements is done by probabilistic rules on layout. The knowledge base also comprises the bibliographic standards and rules. The plans are to make use of authority files to insert standardized information as well.
In the course of the development we are looking for other artificial intelligence components to be possibly integrated.

The technical development will certainly lead us to more decentralized - distributed - processing. This presupposes a so-called 'Mini-aDIS', which, however, will not be available in the near future.

REFERENCES

[1] Bürk, K.; Marek, D.: Produktion von wissenschaftlich-technischen Datenbanken. In: Handbuch der Modernen Datenverarbeitung. Wiesbaden: Forkel-Verl., No. 141. (May 1988)

[2] Niedermeyr, W.; Kappus, H.: Bibliotheksverbundsystem (BVS) und sein Einsatz im Fachinformationszentrum Energie, Physik, Mathematik. In: ABI-Technik 2 (1982), 1. S. 7-15

[3] Marek, D.: Zwei Jahre Online-Input im Fachinformationszentrum Energie, Physik, Mathematik. In: ABI-Technik 3 (1983), 3. S. 201-208

[4] Marek, D.: Die Beschaffung der dokumentarischen Bezugseinheiten. In: Laisiepen, K.; Lutterbeck, E.; Meyer-Uhlenried, K.-H. (Hrsg.): Grundlagen der praktischen Information und Dokumenation. Eine Einführung. 2. Aufl., Saur, München, New York, London, Paris, 1989. S. 192-213

[5] Einführung in das adaptierbare Dokumentations- und Informationssystem aDIS, München: aStec angewandte Systemtechnik GmbH, ca. 1988

[6] Test: Lesesystem Discover 7320. Sonderdruck aus: PC Magazin. München: Markt & Technik Verl. Ausg. 49, 1987

[7] Übersetzungscomputer sind marktreif. Sonderdruck aus: Siemens Zeitschrift. Vol. 62(N. 1), Jan./Feb. 1988

[8] AUTOCAT - Wissensbasiertes Formalerfassungssystem nach INIS-Regeln am Beispiel von Kernzeitschriften des Faches Physik. Sachbericht für den Zeitraum vom 1.1.86 bis 31.12.86. (Bericht DV-II-87-5 (AUTOCAT 87-2)). Darmstadt: Technische Hochschule, 1987

[9] Rauth, U.; Schnellbach, C.: Das AUTOCAT-Konzept der wissensbasierten Formalerfassung von Zeitschriftenaufsätzen. In: Von der Information zum Wissen - vom Wissen zur Information: traditionelle und moderne Informationssysteme für Wiss. u. Praxis/Dt. Dokumentartag 1987. Weinheim: VCH, 1988. S. 303-318

Fig. 1 Architecture of Realtime Applications

Fig. 2 Single Data Capture and Data Storage

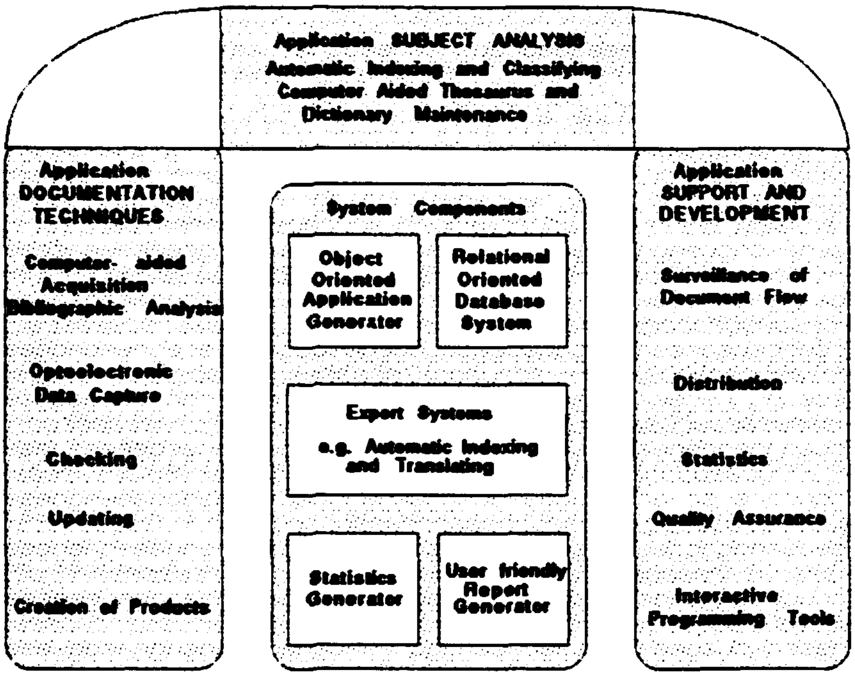
Fig. 3 Nonredundant Storage of a Documentary Unit

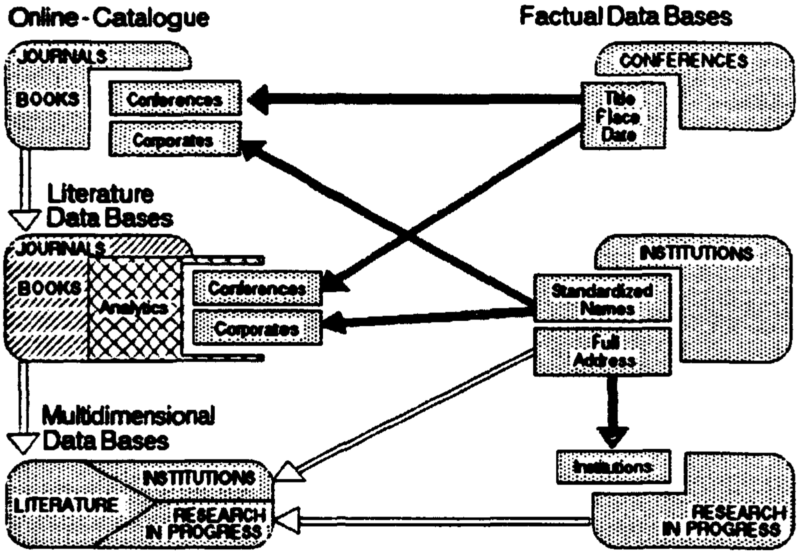
Fig. 4 Optimizing Documentation Process

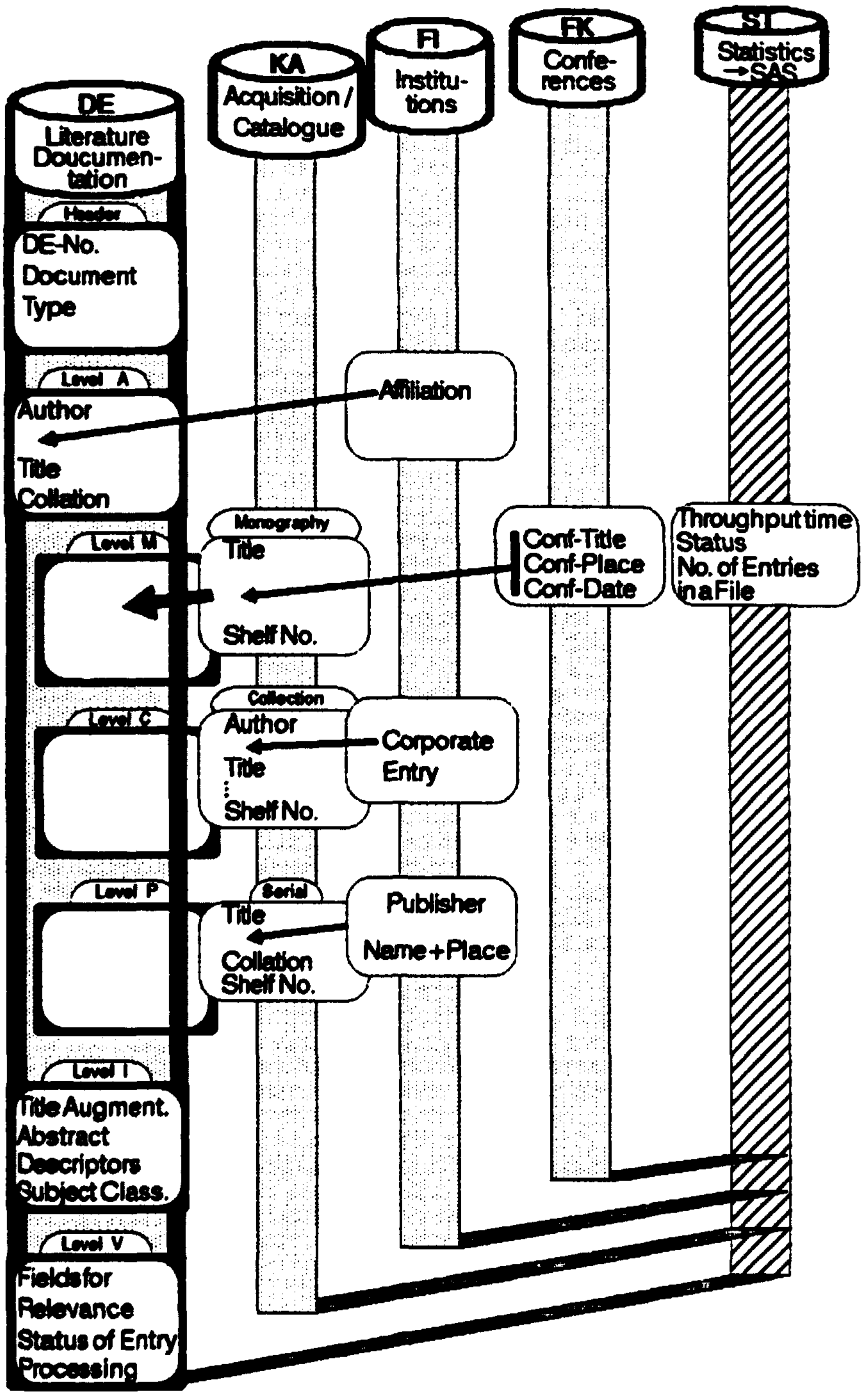
Fig. 5 ADIS Features

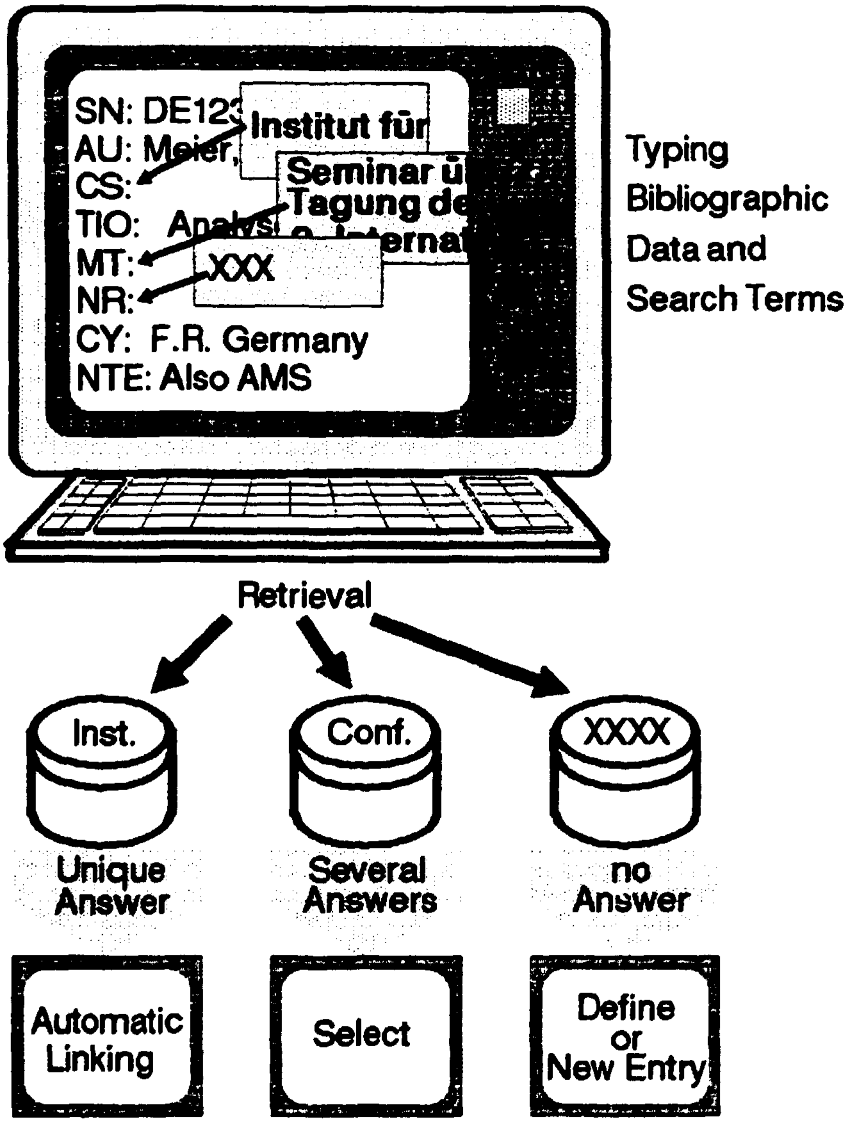
Fig. 6 Thesaurus Construction and Maintenance

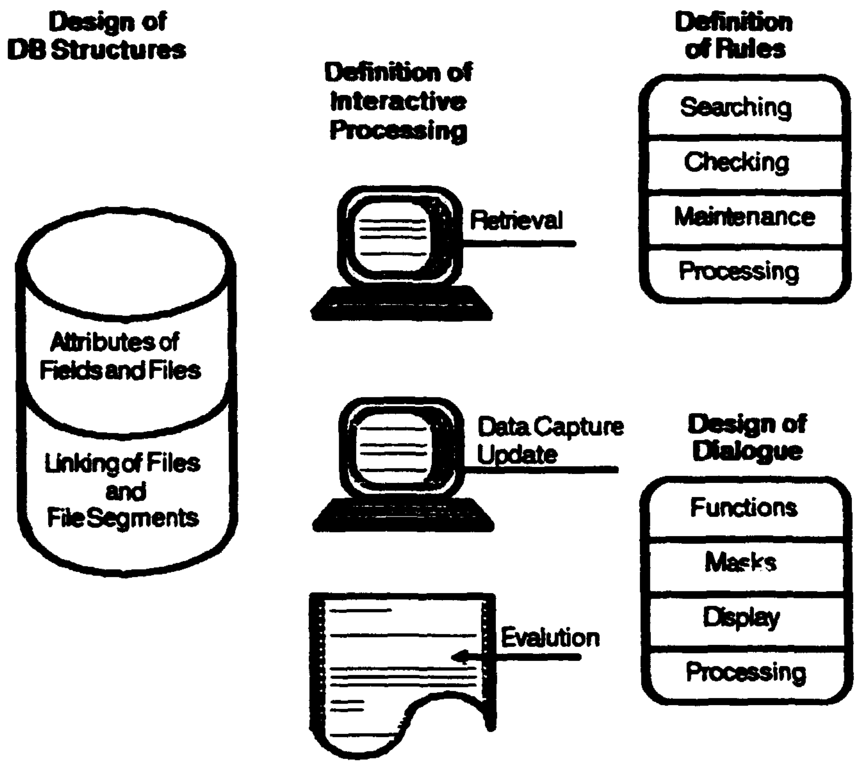
Fig. 7 Computerized Periodicals Management

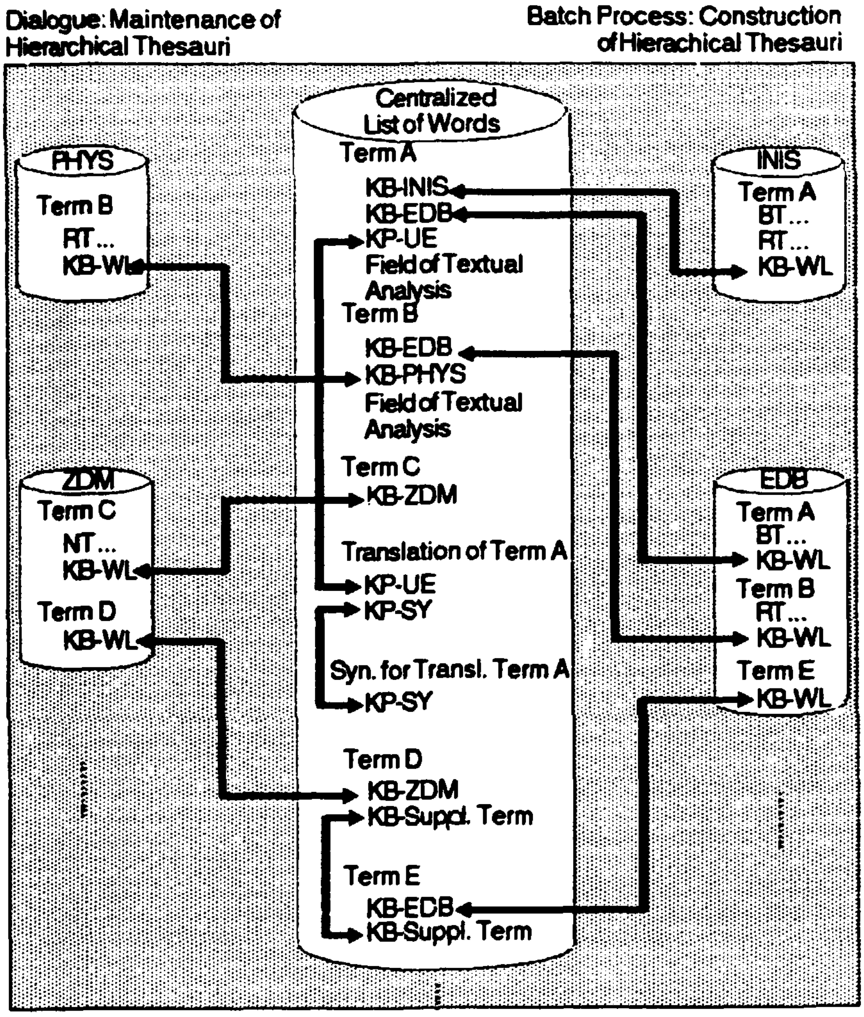

Application SUBJECT ANALYSIS
Automatic Indexing and Classifying
Computer Aided Thesaurus and Dictionary Maintenance
Application DOCUMENTATION TECHNIQUES
Computer- aided Acquisition
Bibliographic Analysis
Optoelectronic Data Capture
Checking
Updating
Creation of Products
System Components
Object Oriented Application Generator
Relational Oriented Database System
Expert Systems
e.g. Automatic Indexing and Translating
Statistics Generator
User friendly Report Generator
Application SUPPORT AND DEVELOPMENT
Surveillance of Document Flow
Distribution
Statistics
Quality Assurance
Interactive Programming Tools

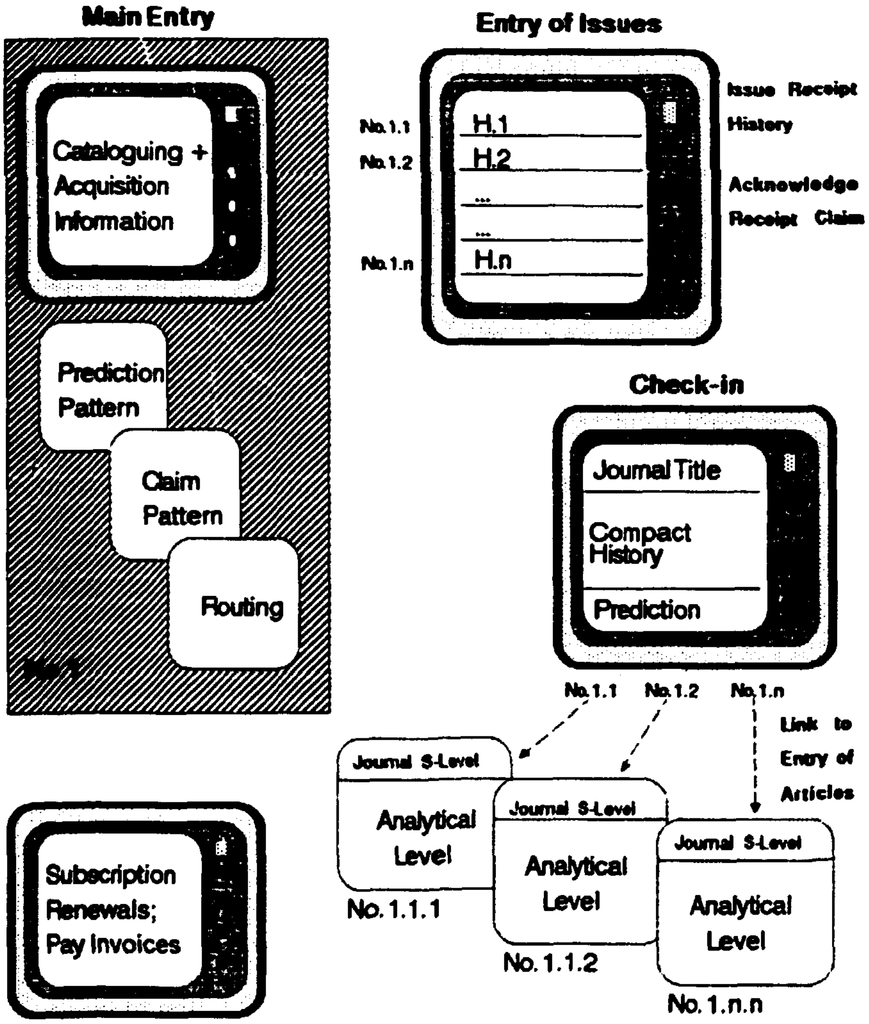

Online - Catalogue
Factual Data Bases
JOURNALS
BOOKS
Conferences
Corporates
CONFERENCES
Title
Place
Date
Literature
Data Bases
JOURNALS
BOOKS
Analytics
Conferences
Corporates
INSTITUTIONS
Standardized
Names
Full
Address
Multidimensional
Data Bases
INSTITUTIONS
LITERATURE
RESEARCH
IN PROGRESS
Institutions
RESEARCH
IN PROGRESS


DE
Literature
Doucumentation
Header
DE-No.
Document
Type
Level A
Author
Title
Collation
Level M
Level C
Level P
Level I
Title Augment.
Abstract
Descriptors
Subject Class.
Level V
Fields for
Relevance
Status of Entry
Processing
KA
Acquisition /
Catalogue
Monography
Title
Shelf No.
Collection
Author
Title
Shelf No.
Serial
Title
Collation
Shelf No.
FI
Institutions
Affiliation
Corporate
Entry
Publisher
Name + Place
FK
Conferences
Conf-Title
Conf-Place
Conf-Date
ST
Statistics
→SAS
Throughput time
Status
No. of Entries
in a File


SN: DE123
AU: Meier,
CS:
TIO: Analys
MT:
NR:
CY: F.R. Germany
NTE: Also AMS
Institut für
Seminar ü
Tagung de
o. Internat
XXX
Typing Bibliographic Data and Search Terms
Retrieval
Inst.
Conf.
XXXX
Unique Answer
Several Answers
no Answer
Automatic Linking
Select
Define or New Entry


Design of DB Structures
Attributes of Fields and Files
Linking of Files and File Segments
Definition of Interactive Processing
Retrieval
Data Capture Update
Evalution
Definition of Rules
Searching
Checking
Maintenance
Processing
Design of Dialogue
Functions
Masks
Display
Processing


Dialogue: Maintenance of Hierarchical Thesauri
Batch Process: Construction of Hierachical Thesauri
Centralized List of Words
Term A
KB-INIS
KB-EDB
KP-UE
Field of Textual Analysis
Term B
KB-EDB
KB-PHYS
Field of Textual Analysis
Term C
KB-ZDM
Translation of Term A
KP-UE
KP-SY
Syn. for Transl. Term A
KP-SY
Term D
KB-ZDM
KB-Suppl. Term
Term E
KB-EDB
KB-Suppl. Term
PHYS
Term B
RT...
KB-WL
ZDM
Term C
NT...
KB-WL
Term D
KB-WL
INIS
Term A
BT...
RT...
KB-WL
EDB
Term A
BT...
KB-WL
Term B
RT...
KB-WL
Term E
KB-WL


Main Entry
Cataloguing + Acquisition Information
Prediction Pattern
Claim Pattern
Routing
Entry of Issues
No.1.1
No.1.2
No.1.n
H.1
H.2
...
...
H.n
Issue Receipt History
Acknowledge Receipt Claim
Check-in
Journal Title
Compact History
Prediction
No.1.1
No.1.2
No.1.n
Link to Entry of Articles
Journal S-Level
Analytical Level
No. 1.1.1
Journal S-Level
Analytical Level
No. 1.1.2
Journal S-Level
Analytical Level
No. 1.n.n
Subscription Renewals; Pay Invoices