BLOCK 4

INFORMATION STORAGE AND RETRIEVAL SYSTEMS

UNIT 14 ISAR SYSTEMS: OBJECTIVES, TYPES, OPERATIONS AND DESIGN

Structure

14.0 OBJECTIVES

In the previous Units of this course we have discussed the basics of information processing and organization. All the areas such as indexing, bibliographic description and formats, vocabulary aspects etc. are the sub-systems of Information Storage and Retrieval System (ISAR). Now we will be taking a holistic view of the total bibliographic information storage and retrieval system. In this Unit, the objectives, types and design aspects of ISAR system will be discussed.

After reading this Unit, you should be able to:

- define an Information Storage and Retrieval System;
- know different type of Information Storage and Retrieval System (ISAR);
- understand knowledge representation aspects of ISAR system designed according to information requirements;
- explain the components of ISAR system; and
- appreciate uses of ISAR system in libraries and information centres.

14.1 INTRODUCTION

Information Storage and Retrieval (ISAR) system deals with three basic aspects:

- Information representation
- Information storage and organisation
- Information access.

One of the best examples of ISAR system is library system, where information is stored, processed organised and retrieved on demand. Information could be stored in a book, audio-video, images and so on. The library and information centres endeavours to organize knowledge available in documentary form. The multi-faceted universe of knowledge is represented in libraries in linear form using some classification scheme. At the time of retrieval, specific aids like cataloguing and indexing are used and meaningful information is retrieved.

A lot of emphasis is given on improving the performance of the system. For that librarians have developed classification schemes for helpful arrangement of documents on the shelf so that retrieval can be facilitated. The tools like library catalogue or indexes are developed and further modified to satisfy the different approaches of users toward information. Automated systems reduce the effective time of users in searching for information which, in effect, further improves the performance of the system. Therefore, tools like Database Management System (DBMS) are used for keeping records of a holding of library, which is known as Online Public Access Catalogue (OPAC).

14.2 USERS AND THEIR INFORMATION NEEDS

Several classes of users consume different kinds of information for different purposes. A user when searches for information, may want either advanced level of information or semi-expert level of information or preliminary level of information according to his information needs. For example, a physician may want advanced level of information on medicine in his area of specialization. But the same person may want preliminary level of information on legal matters and semi-expert level of information on biomedical instruments. So the information needs of a user differ significantly in different situations.

In general, the 'Information Gathering Habit' surveys reveal that there are at least four types of information needs of the users. These may be categorised as:

a) Current Information Need,

b) Exhaustive Information Need,

c) Every day Information Need, and

d) Catching-up or Brushing-up Information Need

You have already learnt about these information needs in Unit 13 of course MLII-101. An ISAR system is designed to meet these types of information needs. Recognition of actual information needs and also potential information needs of the users is a problem faced by the ISAR system developers, as it has been recognized that the 'need' and 'want' do not connote the same meaning. Users may also have different approaches towards of a library. A user might make an approach by:

- name of author
- title
- subject

The system, thus, should have provision of field level search, not only for browsing facility according to author, but also, title, subject and so on. This is quite an imaginative job from the point of view of a system designer. In a typical library automation system, the front-end should have facilities of generating different kind of reports of library usage.

But most commercial ISAR systems cater to either exhaustive information needs or current information needs of the users. The different classes of users can be professionals, students, researchers, policy makers, managers, and common citizens. Many times such information may be available in sources other than the documentary sources. For every profession and every vocation, there is demand of information to successfully carry out professional or vocational work. For example, a farmer may require weather-related information as well as information on best farm practices. A journalist may want background information for his current assignments, but not necessarily from the documentary information sources.

The ISAR system is designed to meet varieties of needs of different classes of users. Nowadays ISAR system is designed on modular basis, which has different components or subsystems. The user interface is designed in a friendly manner (user friendliness), so that ease of use can be ensured. An ISAR system usually offers different approaches of searching, which may be achieved through strong indexing capabilities and facilities. Users also have choice to search through simple search, or advanced search mechanisms. In simple search, user enters a single search parameter such as title of document or keyword or name of author. In advanced search, user enters a combination of search parameters, such as, more than one keyword, or a keyword and a name of author, or, a keyword and year of publication, etc. In advanced search, Boolean operators, like AND, OR, NOT, SAME, etc. can be incorporated to refine a query. The system which is to be developed should have all these capabilities to search and retrieve information.

14.3 OBJECTIVES OF ISAR SYSTEMS

The principle objective of ISAR system is to provide correct information to the user in least time with least efforts. Thus, while designing any ISAR a system designer should keep following objectives in view:

Information Facilitator

The ISAR system should act as facilitator between the information (contained in document) and the users. If a user approaches with the subject term, name of contributors or title of the document and so on, the system should be helpful to give him the desired information. The information could be exact information or the reference of a document which contains information.

Non-Ambiguous

The system should be so organized that ambiguity of information is avoided so that search result is free from any kind of ambiguity. This requires identification of terms, setting their context and their proper indexing. For example, search for a term 'screw driver' should not bring results like 'truck driver', 'hardware driver' and so on.

Minimum Time

The system should be so designed that minimum effort and time are spent to interrogate the system.

Searching through the system should take minimum time, meaning thereby that the ISAR should be capable of performing fast search. Not only that, it is best

to have an online ISAR so that users do not need to walk to library. They should get whatever they want at there work place.

User Friendliness

Ease of use is an important consideration for any ISAR system.
Any ISAR should have user friendly interface. The important aspects of ISAR should be highlighted. Before a user uses the system he/she should be properly introduced to the system with all its features, i.e., informing users about the scope of system, available search options, and most importantly how to perform search with the system. It is only this interface through which a user operates an ISAR system. Take an example of a Library OPAC. It should have following features:

* Introduction to library
* Scope of collection
* Instructions for performing search

The search interface should facilitate framing the search like,

* Keyword search
* Author and title search
* Combination search (using Boolean operators)
* Proximity search, etc.

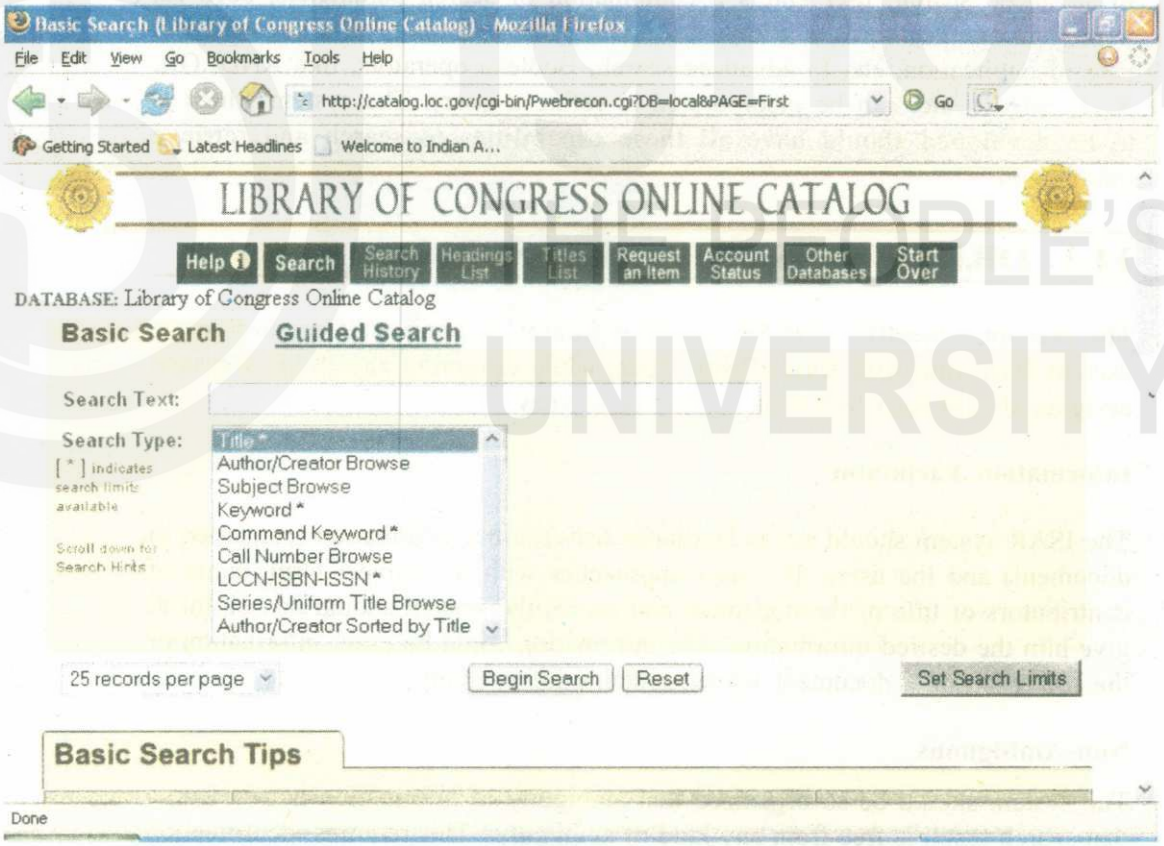


Fig. 14.1: Web OPAC of Library of Congress

While designing an ISAR system the following aspects should be recognised to achieve the objectives of the system:

* The desirability of making systems as readily usable as possible for their clienteles;
* The need to recognise basic features of retrieval system; and

- To incorporate coordinating features such as vocabulary control, search strategies, user-interface, information modelling aspects in general, etc.

Features of ISAR System

Keeping in view the objectives mentioned above and recognizing the aspects to be considered in designing a system, an ideal ISAR system should incorporate one or more of the following features:

- The competence and compatibility for consolidated searching and retrieval of information from any client terminal from any database within the system.
- It should be able to narrowcast or broadcast or relate the information need in a variety of associations to get optimum retrieval performance.
- It should have access facilities at multi-points.
- It should have common command language facility to retrieve information from several databases of the system.
- It should be able to handle information access from entity-related or object-oriented approaches. It may also provide all other associations for accessing information.
- In a bibliographic or full-text database, the surrogates chosen should have indicative as well as informative features that are sufficient enough to select or reject the retrieving information based on end-users' needs.
- It should have the ability to select, classify, process and consolidate the analysed information into a cohesive text ready for assimilation by the end-users.
- It should have ability to orient the information to specialist needs of the users from time to time. This calls for understanding the processing of user profiles.
- It should be able to retrieve maximum information with minimum number of clues. The fuzzy approaches of end-users must be able to get clarified and ultimate result should provide satisfaction to the searcher.
- It should have capacity to interchange the information available in one database or another for purposes of retrieval relevance end usage.
- It should have bibliographic data interchange capacity (using Z39.50 or similar standard) to meet consolidation to a chosen format for networking and other purposes. Compatibility with standards at all levels must be the goal.
- It should have ability to search simple information quickly in an easy manner and also have the ability to multi-track the complex questions and present them in a simple easy manner. User-friendly presentations are very important.

14.4 TYPES OF ISAR SYSTEMS

ISAR systems are used by a wide range of users. According to different kinds of needs and purpose of use, different types of automated systems may be designed. Such types may be:

- Database Management System (DBMS)
- Text Retrieval System
- Management Information System (MIS)
- Decision Support System (DSS)
- Knowledge Based System (KBS)

14.4.1 Database Management System (DBMS)

Any automated system is based on a collection of stored information or documents

in a database which is amenable for access. A DBMS is primarily concerned with data storage, maintenance and retrieval and is used to keep control and manipulate data within the database.

The distinguishing characteristic in DBMS is the definite structure of the stored information, instead of dealing with natural language text. In DBMS, normally files of data are described by a small set of pre-specified attributes. For example, in a Salary database of an organization, name of person, designation, salary, etc. are attributes. Similarly, in the context of records of books author, title, publisher, year etc., may be the attributes. Each attribute carries some kind of value in it. Therefore, a DBMS can be defined as set of records and each record contains fields (attributes) which in turn contain data (value). A database may contain textual, numeric, statistical and graphical information.

OPAC is a kind of DBMS often built of some kind of Bibliographic Database Management Systems (BDBMS). The typical example of BDBMS is the one built by CDS-ISIS/WINISIS developed by UNESCO. The data fields may contain author, title, place, publisher, year of publication and so on.

14.4.2 Text Retrieval System

In contrast to DBMS, text retrieval systems are designed for unstructured data such as full text documents. Queries are usually language based here such as, keywords and a number of advanced search techniques (such as proximity search) can be used. However, systems may also handle discrete structured data.

14.4.3 Management Information System (MIS)

Management Information System is a kind of database management system designed to cope up with the needs of managements who need to have information about different alternatives related to his/her interest to facilitate his work. Though built on DBMS platform, information are subjected to special processing. In such a system information is available with different alternatives. In the business environment managers needs to take complex and rapid decisions. Under such circumstance MIS is a kind of sophisticated tool which provides them timely information to take decisions.

Collection of data is critical in MIS because the information comes from different sources, i.e., within the organization or from outside organization. Collection of data not only needs defining that how the data would be captured but also the estimation of cost involved in data collection. Once the data are collected and organized such system generates reports for usage. The reports could be generated in printed or electronic form depending on desired format. Such systems also generate reports upon different intervals if it is desired.

14.4.4 Decision Support System (DSS)

Decision Support systems help top-level management in arriving at decisions. There is very little difference between MIS and DSS. The former generates reports in anticipation or on demand and collect facts, whereas the later provides possible alternative solutions. These are interactive computer-based systems that provide the user easy access to decision models and data in order to support semi-structured and unstructured decisions. In management parlance a structured decision means use of rules and norms for making decision. Such systems help managers in identifying the problem, analyzing alternatives and choosing possible solution. However, the typical decision-making takes place with shared effort of human and machine. In true sense, a DSS cannot take decision, rather it amplifies decision maker's capability by providing resources and facilities. In other words it provides intellectual support. These systems are integrated with powerful tools like generating charts, preparing tables and presentation tools.

14.4.5 Knowledge Based System (KBS)

Specialised computer programs, modeled in the same way as human experts tackle problems and arrive at solutions, are called 'Expert Systems'. Such systems rely upon a store of specialised knowledge for solving problems and hence referred to as Knowledge Based Computer Systems (KBCS) or Knowledge Based systems (KBS).

Expert systems are sophisticated computer programs that manipulate knowledge to solve problems efficiently in a narrow problem area. Knowledge based systems enhance the value of expert knowledge by making it readily and widely accessible. Like human experts, these systems use symbolic logic and heuristics to find solutions. They are also capable of learning from experience through inferencing mechanism.

KBS systems are domain specific and are backed up by a strong knowledge base. In these systems each bit of information is not only stored but they are also linked. This linking is used to preserve the context. The context is used to draw the inference from a query. They are capable of providing solutions and replace the human intervention. Expert systems are thus data driven. But it is also important that how the given information is utilized to achieve the goal. Human experts possess procedural knowledge (How To), which helps them to flowchart the courses of action to be taken in solving problems. Accordingly, rules for manipulation of the knowledge have to be incorporated in the expert systems to get the desirable results. However, this procedure involves other related tasks such as intuitive inferencing power, learning and updation of knowledge. Solutions to problems can be achieved if only all tasks are executed in a coordinated way.

Self Check Exercises

1) What are the objectives which should be kept in mind while designing an ISAR?

2) Name different types of ISAR systems.

Note: i) Write your answers in the space given below.
ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

14.5 DESIGN OF ISAR SYSTEMS

From the foregoing discussion, it should be clear that a number of features should be ingrained in an ISAR system. Thus, an ISAR system is an integrated one combining various aspects.

14.5.1 Components of An ISAR System

An ISAR system comprises of following components:

- User Interface
- Knowledge Base
- Search Agent

The features to be incorporated in each components are discussed below :

14.5.1.1 User Interface

User interface is the part which puts users across the ISAR system. It is the front-end which enables user to put a query and displays results. Basically, user interface is of two types:

- Query Interface
- Result Interface

14.5.1.1.1 Query Interface

This is the end from where users enter his/her search terms. It is one of the major components which initiate communication between users and the system. The Query Interface should have following features:

a) *Understanding the user input statement*

It is also known as front-end. The interface should be able to capture keywords given by users which should be passed on to the search program. The front-end should have understandable look and feel. At the time of designing the system, one should also seriously consider use of different colours, and instructions for performing search and limitation of search.

b) *Refining the problem statement*

Sometimes users start with broader domain and further refine the search. The interface should have flexibility for further refining any query within the displayed search results. In other words, system should provide facilities for further modification of search statement. It should also display some kind of arrangement among topical terms which further facilitate browsing through the system.

c) *Search statement to search strategy translation*

Any knowledge base accepts a query in a particular format. For example, a Relational Database Management System (RDBMS) accepts search statement in Structured Query Language (SQL). It is the system front-end which translates a search statement and formulates a search strategy in the language which is understood by Search Agent.

d) *Modification of search strategy*

If one does not get desired output from the database, ISAR system should have procedure for further modification of search strategy. The modification should be interactive. Vocabulary control devices can also be added as an aid for users to locate the term of his/her interest.

14.5.1.1.2 Result Interface

Display of search results is another important aspect of searching. It should be in user friendly manner. Not only that the result should cater the needs of individual users but the display should also be customized. Search results should also display the ratings in the light of search terms. For this purpose statistical techniques can be used.

14.5.1.2 Knowledge Base

The store house of any ISAR system is its Knowledge Base. It contains list of facts or related facts (information). Any kind of query is answered based on the facts stored in the Knowledge Base. A Knowledge Base could be a Database Management System (DBMS). Retrieval of information from storage depends on two important aspects of Knowledge Base:

- Knowledge Representation
- Indexing and clustering

14.5.1.2.1 Knowledge Representation

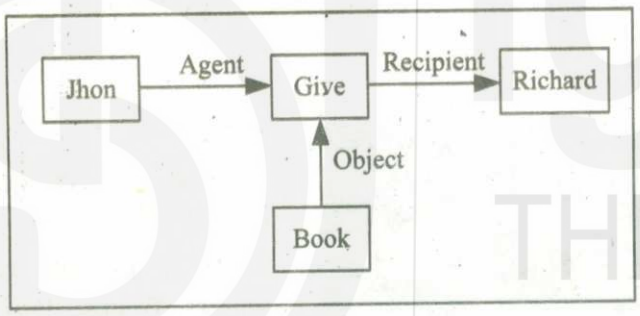
The first and foremost objective in constructing an ISAR system is representation of facts within the Knowledge Base. There are different ways of representation of knowledge:

Semantic Network Knowledge Representation

Semantic network is a method of knowledge representation based on a network structure. A semantic network contains points called *nodes* connected by links called as *arcs*. The nodes represent objects, concepts or events - in other words documents or information. The arcs are used to represent the relations between the nodes. Arcs build a kind of hierarchies in the Knowledge Base. Arcs usually represent relations like *is_a* or *has_part*. For example, Universe of Knowledge à Library Science à Cataloguing

Semantic networks are a useful way to represent knowledge in domains that use well established taxonomies to simplify problem solving. Semantic networks are useful in representation of sentences of natural language.

For example, a sentence *John gave a book to Richard* can be represented as,



Frame Based Knowledge Representation

It is an object-oriented approach. A frame represents an object (document or information) or class of objects (collection of documents or information) or several facts. When they represent a class of objects, they generalise certain groups identifying overall properties of those groups, it shares. The pointers where properties are stored are known as *slots*. Similarly, if frame represents an object, slots represent the properties or attributes of the object. Slots contain value for that particular attribute. For example, a book in a library is an object, therefore it can be represented as *frame*. The properties of book, i.e., Title, Author, Place, Publisher and so on are stored as slots and each slot would have corresponding value.

Frames are also very helpful in representing hierarchies. In a frame base, slots can store procedures or relations. Relations are used for storing taxonomy or genealogical data or parent child information. For example, if a person is represented by a frame, different slots can store value who is the father or who is the son of the person. Similarly, slots can also store procedures that means how a frame or object should react or the way it should be used or it can operate with other objects.

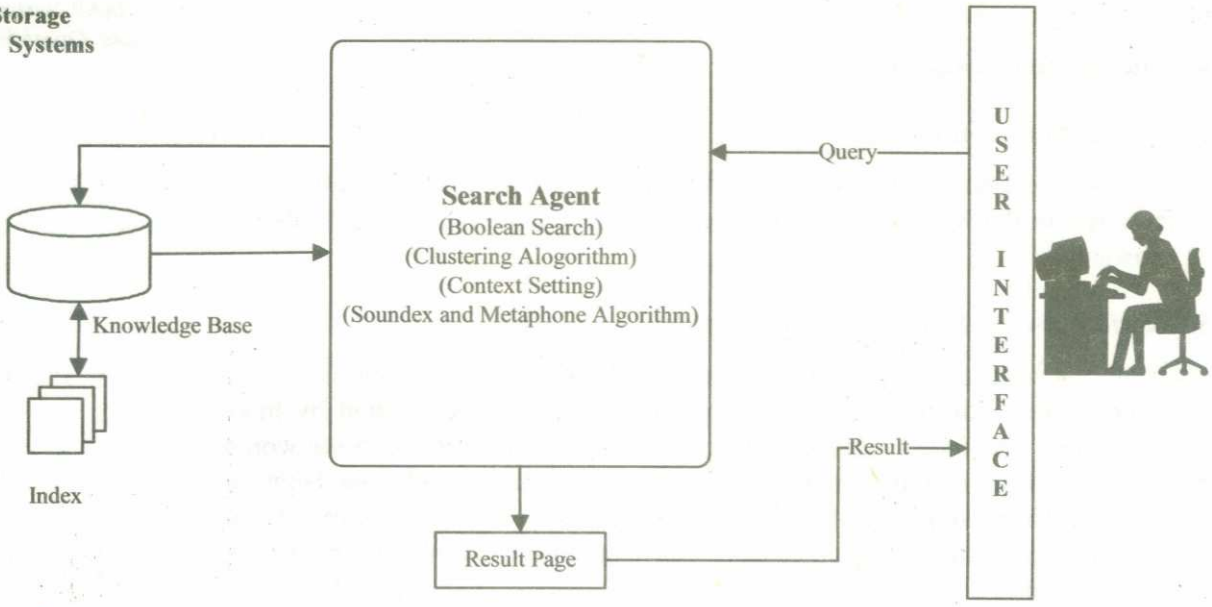


Fig. 14.2: Diagrammatical representation of an information storage and retrieval system

Rule-Based Knowledge Representation

Rule based representation is a popular approach. Rules are employed to state the way in which the inferencing has to be done. Rules provide a formal way of representing recommendations, directives, or strategies. Rules are appropriate when the domain knowledge results from empirical associations developed through years of experience in solving problems in a given area. Rules are expressed in the form of IF-THEN statements. For example,

If search is in collection of BOOKS	THEN display Title, Author, Place, Publisher, Year, Physical Description, ISBN
If search is in collection of ARTICLES	THEN display Title, Author, Name of Journal, Volume, Issue, Year, ISSN

In a rule based expert system, the domain knowledge is represented as a set of rules that are checked against a collection of facts or knowledge about the current situation. When the IF portion of the rule is satisfied by the facts, the action specified by the THEN portion is performed. When the condition is satisfied the rule is said to 'fire' or 'execute'. A rule interpreter is used to compare the IF portions of rules with the facts and execute the rule whose IF portion matches the facts.

Indexing

Many of the systems follow a kind of keyword indexing. Unfortunately the keyword indexes are sometimes good in recall not in precision and in some cases vice-versa. Such indexes fail in preserving the context of search term. WWW is a good example of use of keyword indexing. There has been different efforts made in this direction such as using clustering techniques like keyword clustering to attach semantics with a keyword. In such a technique, relation among the terms are used like Broader term, Narrower term and Related term. This technique heavily uses Thesaurus. Such relations can be stored in the form of sequential inverted files or using B-tree structures.

14.5.1.3 Search Agent

Any ISAR system should be backed up by a Search Agent. It is a program which takes input from Search Interface and searches in the Knowledge Base

using existing index. A good ISAR system means efficient retrieval. Thus, a good search agent must be equipped with following features:

- facility of using Boolean operators
- context setting to search terms
- use of clustering algorithms
- use of soundex and metaphone algorithms

14.5.1.3.1 Boolean Operators

Three Boolean operators are AND, OR and NOT. These operators are used to generate combinational search. AND and NOT operators increase precision whereas OR increases recall of search results. The shaded area represents retrieved records in the following example.

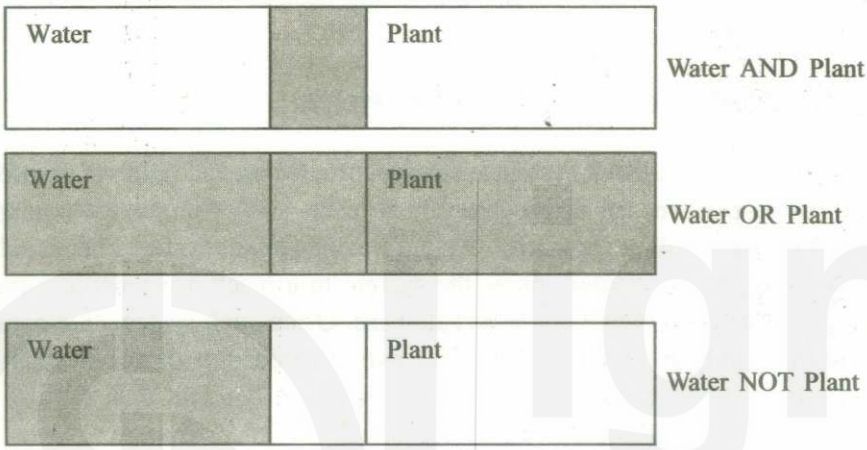


Fig. 14.3: Boolean operators

14.5.1.3.2 Context Setting

Context Setting requires content analysis of document. Here one analyses document manually or automatically in order to preserve the context of each term in the index. It can be done in two ways, i.e., Conceptual Analysis and Relational Analysis. Conceptual analysis can be thought of as *frequency of concepts*. Concept can be represented by texts as well as pictures where text is very common. To analyze the concept one looks for the appearance of words in the text. It is not necessary that same word appears always, there may be synonymous terms present. For example, if one is analyzing with a hypothesis that a certain document is about *freedom* then one should look for the related words like *liberation*, *independence*, etc. In contrast, relational analysis goes one step further by examining the relationships among concepts in a text. In relational analysis we look for what are the related words appearing next to the word in question. For example, to see what are the words that appear next to *freedom* and then determine the related concepts.

14.5.1.3.3 Clustering Algorithms

Clustering is a method by which large sets of data is grouped into groups or clusters of smaller sets of similar data based on some characteristics. For example, in a group of players one can cluster players according to their specialisation of game, like those who play cricket, those who play hockey and so on. A clustering algorithm attempts to identify natural groups of components or data based on some similarity in a given population. In other words, it is a method to create sub-class in a given class. The first thing in such algorithms are identification of core entity which is also known as *centroid*. Around centroid similar kind of entities

or data are collected which are called as members of cluster. To determine cluster membership and size, most algorithms evaluate the distance between each entity or data and the cluster centroids. Statistical techniques are used for generating clusters as output.

14.5.1.3.4 Soundex and Metaphone Algorithms

Soundex and metaphone algorithms are almost the same kind of algorithm. Both these algorithms are based in the way pronunciation of a word is made. In soundex algorithm, a numeric code is assigned to each character used in a word and when search is performed, words with similar codes are also brought out in search result. Metaphone is also same kind of algorithm but unlike soundex which encodes on letter-by-letter basis, it encodes groups of letters i.e. a word. Metaphone embodies more accurately the rules of pronunciation in language. Such algorithms are well established for English as a language. Both algorithms return all the words that exactly match the desired word as well as all similar sounding names.

14.5.2 Functionality of An ISAR System

The objective of an ISAR system is storage, processing and retrieval of information. In order to perform the above mentioned tasks, such systems are equipped with user friendly interface, powerful database management system and search agents. Information is represented inside the system in a machine readable format using frames or semantic networks or as rule base. Often such systems are combination of above three or two methods. The overall objective of the system depends upon the need of users. Once needs of users are defined, it is easy to frame a system for the purpose.

System stores information in a form of structured record which in turn is indexed using automated or manual technique. Often the search engines over Internet use Keyword indexes using automated programs called Robots or Spiders. It goes through various web pages on Internet and indexes them word-by-word. This index is stored inside databases similar to library catalogues. Any kind of search query is searched through the index created by Robots.

From the perspective of users, he/she approaches to ISAR system with some query in his/her mind. Then he/she transfers the query into a search strategy with fixed vocabulary or key term or in a fixed format (for example, SQL query). Once query is made, he/she is given a list of search results out of which one chooses the best ones of his/her interest.

At this level a user goes through the search results and evaluates the displayed result in the light of his/her query. If the result displayed does not suit to his/her needs then he/she can go for further refinement of search strategy.

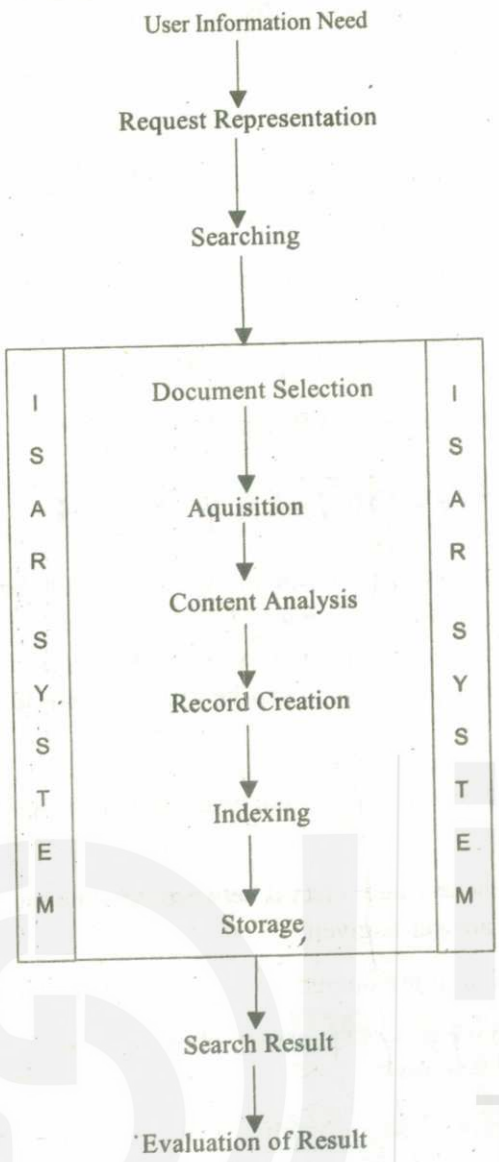


Fig. 14.4: Functional view of ISAR

Self Check Exercise

3) What are the different components of ISAR system?

Note: i) Write your answer in the space given below.
ii) Check your answer with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

Steps in the Development of ISAR Systems

The development of an ISAR system starts with the recognition of the need for developing a system in relation to the needs of the users, as all the subsequent activities are dependent upon these. When designing, ISAR systems should follow system development life cycle (SDLC) for greater efficiency and effectiveness of the systems. In the design of the systems, following steps need to be followed :

a) Recognising the need for development of ISAR system.

b) Recognising the information needs of the users.

c) Identification of users need.

d) Deciding the type(s) of databases to be incorporated into the system.

e) Deciding about the features for incorporation in the databases.

f) Preparation of structured queries.

g) Design and development of various components of the system such as user interface, search agent, etc.

h) Evaluation of the system.

i) Re-designing/Modification of ISAR system, if needed.

14.6 EVALUATION OF ISAR SYSTEMS

In an ISAR system search can be performed in manual or automated environment. An example of manual system is library shelf-list or card catalogue whereas OPAC or Digital Libraries are examples of automated systems. According to Claverdon and others any ISAR system can be evaluated based on the following points:

- The coverage of the collection, that is, the extent to which the system includes relevant matter;
- The time lag, that is, the average interval between the time the search request is made and the time an answer is given;
- The form of presentation of the output;
- The effort involved on the part of the user in obtaining answers to his search requests;
- The recall of the system, that is, the proportion of relevant material actually retrieved in answer to a search request; and
- The precision of the system, that is, the proportion of retrieved material that is actually relevant.

If the view of Claverdon is taken and further analyzed one gets following check points, on which any ISAR can be evaluated:

- **Coverage**
- **Cost Benefit Analysis**
- **Time**

The important issue which should be taken into consideration while evaluating an ISAR, is the Time factor. The first and foremost thing which should be observed, How long a user takes to get the satisfactory answer to his question? If a user spends more time over system searching for his/her answer, it means that the system has to be modified in such a way that one gets his/her answer immediately. Searching over an ISAR system depends a lot and varies from on person to person. Therefore such kind of measurement should be supported by a long observation.

Other evaluation criteria are:

- **Completeness and Relevance**

- **Novelty Ratio**
- **Noise**

These have been discussed in detail in Unit 5 of this Course. The evaluation parameters discussed therein in the context of indexing systems are equally applicable for ISAR System as a whole or a sub-system of it.

Critical Aspects of Recall and Precision

There is a problem of determination of the total number of relevant documents in collections unless one scans the whole collections completely, which is almost unfeasible. Another thing is that relevancy cannot be measured only by yes/no terms, but also by the degree of relevancy. Recall and precision have inverse-relationship, where both cannot achieve the highest degree at a point of time. The recall and precision are influenced by the following factors:

- queries that imperfectly represent information needs;
- indexing factors;
- search strategy factors; and
- vocabulary factors.

Lack of specificity, lack of exhaustivity, lack of specific terms, inadequate hierarchical cross-reference structure, defects in hierarchy, failure to cover all reasonable approaches to retrieval, etc. are the major hindrance to achieve higher degree of relevant retrieval from an ISAR system which also affect recall and precision ratios.

14.6.1 Systems Criteria for Evaluation

The basis for evaluation of an ISAR system could be done at every stage with every component of it. Once again it centres around the users, information sources, intermediaries, the tools, techniques, methodologies and overall environment. B.C. Vickery and Alina Vickery have provided a detailed structure for it.

After setting a framework for evaluation of ISAR system, they present a panorama and yet analyse view of ISAR system which is on the basis of quality and value as beneficial aspects of the evaluation system. User demands and their satisfaction through provision of right information will be ultimate aim of any evaluation. They have structured their thoughts into following aspects:

Table 14.1: Criteria for Evaluation

S.No.	ASPECTS	PARAMETERS
1.	Criteria for Evaluation	Quality and value
2.	A Frame Work of Evaluation	Correlation of Information Provision, Information Content and Information use
3.	Relevance and Assessment	Source and Receiver coexistence
4.	Service Qualities	Simplicity, ease of use, personal attention, internal decor and presentability, success/ failure analysis
5.	Evaluating Performance	Choice of characteristics for Performance. Measure of Performance
6.	System Efficiency, Cost, Manpower and Cost-Effectiveness	Labour, Expenditure on documents consumables, Equipment External Charges, Service overhead management and development
7.	Coverage in Acquisition Search	Messages acquired/ Messages omitted
8.	Retrieval from Store	Relevant/ Not relevant, Precision of retrieval
9.	Evaluation of an Information System	Study of Hits/ Non Hits Relevance of Hits Search Qualities
10.	Operational Current Awareness Service	User assessment of processing, indexing and people studies
11.	Online Search Service	Value of computer-based information searches Value of Internet-based information searches
12.	Experimental Study of Retrieval	Intellectual analysis and its role on operations in retrieval
13.	Availability on Demand	Compatibility between Database and the User's demands
14.	Variables affecting Availability	Information and studies
15.	Document-delivery Test	Actual demand potential demand time. Time lag studies.
16.	The Effect of Service Delay	Impact on access Impact on queuing Impact on service
17.	Degradation of Performance	Queue imputes Set priorities in the queries
18.	Value of Information	Reference Introduce self-service personal knowledge Synthesis Measurement of Relevance
19.	The Perceived Value of Information Services	Matrix of value for professional literature

Nowadays advanced technologies are used in modern ISAR systems. The full-text and multi-media contents are now common in modern ISAR systems. Also hybrid systems coexist that provide access to physical collections as well as electronic collections. The criteria for evaluation, as suggested by B.C. Vickery and Alina Vickery, are still relevant in modern ISAR systems, but require some modifications to cope up with new techniques, new technologies and new types of documents.

14.7 EXAMPLES OF ISAR SYSTEMS

14.7.1 Online Public Access Catalogue (OPAC)

Library catalogue has changed its form to the electronic form with the advent

of computerisation and today it is called the OPAC (Online Public Access Catalogue). OPAC provides access to the documents by different approaches of users, such as Author, Collaborator, Title, Subject, Keyword, etc. An OPAC is built on Data Base Management System.

An OPAC should provide facilities for searches like Boolean, truncation and proximity searches so that users can locate document without much effort. OPAC uses bibliographic standards like MARC21, CCF (Common Communication Format), UNIMARC etc.

14.7.2 Digital Library

Online Public Access Catalogues (OPAC) provides only bibliographic details not full-text information. This limitation led to development of full-text databases in digital form. This kind of database is known as Digital Library. Digital libraries are available over network often using WWW.

The costs of creating, storing, and transmitting digital information have been decreased, and the technology to support distribution and access is widespread. Rising acquisition and subscription fees (not to mention shelving and processing costs) have forced libraries to seek other ways to make information available.

Perhaps more importantly, digital libraries support service improvement. Information search and navigation across electronic information resources are faster, with enriched points of access and alternative methods for browsing and exploration. The resources themselves can be segmented, rearranged, annotated, and enhanced in ways not possible before. Digital libraries provide remote access to information resources to the users on their desktops. Also, digitization of documents presents opportunities for long-term preservation of bodies of knowledge, if not of the original carriers of that knowledge.

14.7.3 Search Engines

WWW is itself is a big source of information. Almost everything and anything can be found over Internet. Search Engines provide a kind of interface for users to search the web. A Search Engine basically has three components:

- a Robot or Web Crawler
- a Database
- an Agent

Web Crawler goes to each and every site over Internet and indexes each word present in the page or sometimes few lines from the page. This index is stored in search engines database with corresponding URL (Uniform Resource Locator). When a search query is given it searches in databases of search engine and result is generated.

Search Engines can be categorised into three main types:

- Individual Search Engines – text or image based search engine, for example, Google (www.google.com)
- Subject Directories – subject-based search engines, for example, Yahoo! (www.yahoo.com)
- Meta-search Engines – search engine of search engines, for example, Askjeeves (www.askjeeves.com)

14.8 SUMMARY

Libraries are store house of information. In a big collection, it is very difficult to locate a document or any piece of information in the library. For this purpose libraries had evolved tools like catalogue and index. With the advent of Information Technology and development of automation of libraries the tools like catalogue and index are being transformed into Automated Information Storage and Retrieval (ISAR) Systems. A typical ISAR system has following components:

- User interface
- Knowledge base
- Search agent

In course of time and with the expansion of IT, several applications of ISAR has been developed. The most important product of automated ISAR system is Digital Libraries.

14.9 ANSWERS TO SELF CHECK EXERCISES

1) A system designer should keep in mind following points while designing an ISAR:
   - The ISAR should act as facilitator between the information (contained in document) and the users
   - Search results must be free from any kind of ambiguity
   - Searching through the system should take minimum time
   - ISAR should have user friendly interface.

2) There are different kinds of ISAR systems such as:
   - Database Management System (DBMS)
   - Text Retrieval System
   - Management Information System (MIS)
   - Decision Support System
   - Knowledge Based System.

3) An ISAR system has following components:
   - User Interface
   - Knowledge Base
   - Search Agent.

14.10 KEYWORDS

Boolean Query : A query that is a Boolean combination of terms. For example, *Information AND Retrieval, Vision OR Sight.*

Classification : The process of deciding the appropriate category for a given document.

Collection : A group of documents that a user wishes to get information from.

Information Retrieval : The study of systems for indexing, searching, and recalling data, particularly text or other unstructured forms.

Inverted File : A representation for a collection that is essentially an index.

Precision : A standard measure of IR performance, precision is defined as the number of relevant documents retrieved divided by the total number of documents retrieved.

Query : A string of words that characterizes the information that the user seeks.

Recall : A standard measure of IR performance, recall is defined as the number of relevant documents retrieved divided by the total number of relevant documents in the collection.

Relevance : An abstract measure of how well a document satisfies the user's information need. Ideally, system should retrieve all of the relevant documents.

Robot : Any browser program which follows hypertext links and accesses web pages but is not directly under human control. Examples are the search engine spiders, the 'harvesting' programs which extract e-mail addresses and other data from web pages and various intelligent web searching programs.

Spider : Also called a robot, a program that scans the web looking for URLs. It is started at a particular web page, and then access all the links from it. In this manner, it traverses the graph formed by the WWW. It can record information about those servers for the creation of an index or search facility. Most search engines are created using spiders. The problem with them is, if not written properly, they can make a large number of hits on a server in a short space of time, causing the system's performance to decay.

URL (Uniform Resource Locator) : A URL (Uniform Resource Locator) is the address of a file (resource) accessible on the Internet, e.g., http://www.dell.com

14.11 REFERENCES AND FURTHER READING

Bailey, Charles W. (1990). The intelligent reference information system. *CD-ROM Librarian,* 5(8), 10–19.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Massachusetts: MIT Press.

Chowdhury, G.G. (1994). *Information retrieval system*. Calcutta: IASLIC.

Chowdhury, G.G. (1994). *Introduction to modern information retrieval system*. 2nd Edition. London: Facet Publishing.

Cleverdon, C.W. (1970). Progress in documentation: evaluation of information retrieval systems. *Journal of Documentation*, 26, 55-67.

Cleverdon, C.W. (1972). On the inverse relationship of recall and precision. *Journal of Documentation*, 28, 195-201.

Cleverdon, C.W., Mills, J. and Keen, M. (1966). *Factors determining the perfor-*

mance of indexing systems, Vol. 1, Design, Vol.II, Test Results, ASLIB Cranfield Project, Cranfield.

Doyle, L.B. (1965). Is automatic classification a reasonable application of statistical analysis of text?. *Journal of the ACM*, 12, 473-489.

Doyle, L.B. (1975). *Information retrieval and processing*. Los Angeles, California: Melville Publishing.

Ghosh, S.B. and Satpathi, J.N. (eds.) (1998). *Subject indexing systems: concepts, methods and techniques*. Calcutta: Indian Association of Special Libraries and Information Centres (IASLIC)

Gopinath, M.A. (1999). *Evaluation of information storage and retrieval (ISAR) systems*. In: MLIS-03, Block 3, Unit 12 course material. New Delhi: Indira Gandhi National Open University.

Gopinath, M.A. (1999). *Objectives of information storage and retrieval system*. In: MLIS-03, Block 3, Units 9 course materials. New Delhi: Indira Gandhi National Open University.

Knowledge representation in computer – supported environment. <http://www.uia.org/knowledg/knowrep.htm>.

Knowledge representation. <http://blackcat.brynmawr.edu/~dkumar/UGAI/kr.html>.

Lancaster, F.W. (1991). *Indexing and abstracting: theory and practice*. London: Library Association.

Lancaster, F.W. (1968). *Information retrieval systems: characteristics, testing and evaluation*. New York: Wiley.

Leibowitz, J. (1990). *Dynamics of decision support systems and expert systems*. Chicago: Dryden.

Maron, M.E. and Kuhns, J.L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7, 216-244.

Methods of knowledge representation. <http://busfa.vu.edu.au/awenn/kbs/week3/KnowRep.html>.

Minsky, M. (1968). *Semantic information processing*. Massachusetts: MIT Press.

Minsky, M. A. (1975). *Frame for representing knowledge*. In: *Psychology of Computer Vision*. New York: McGraw–Hill.

Paice, C.D. (1977). *Information retrieval and the computer*. London: Macdonald and Jane's.

Refdesk search engine. <http://www.refdesk.com/newsrch.html>.

Salton, G and McGill, M. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Senko, M.E. (1969). *Information storage and retrieval systems*. In: *Advances in Information Systems Science*. New York: Plenum Press.

Sparck Jones, K. (1971). *Automatic keyword classification for information retrieval*. London: Butterworth.

Vickery, B.C. (1970). *Techniques of information retrieval*. London: Butterworth.

Waterman, D.A. (1986). *A guide to expert systems*. Massachusetts: Addison-Wesley.

Weil, C.B. (1968). Automatic retrieval of biographical reference books. *Journal of Library Automation*, 1(4), 239–249.

Zarri, Gian Piero. (1985). Interactive information retrieval: an artificial intelligence approach to deal with bibliographic data. *Advances in Intelligent Information Retrieval: I* Informatics, 8, 101–149.

UNIT 15 COMPATIBILITY OF ISAR SYSTEMS

Structure

15.0 OBJECTIVES

The Information Storage and Retrieval Systems (ISARS) are developed to satisfy users information needs. Though the objectives vary from system to system, there are many features which are common to them. Thus, if there are compatibilities in various aspects, it becomes easier to exchange the resources, thus saving efforts, costs and duplication. It also facilitates users accessibility to information. It is, thus, worthwhile to study the various compatibility issues, methods and their usefulness.

After going through this Unit, you will be able to:

- explain compatibilities of ISARS;
- describe the areas for compatibility amongst ISAR systems;
- outline multi-task approach to compatibility from various components;
- interpret the interactive planning approach for compatibility; and
- explain the quality control aspects in compatibility issues.

15.1 INTRODUCTION

The term Information Storage and Retrieval System (ISARS) in its widest sense covers all types of organisations concerned with the collection and dissemination of information.

The main purpose of compatibility is to facilitate exchange and sharing of resources. The sharing of resources is a concept which is put to action since a long time which covers various activities undertaken jointly by a group of libraries or other organisations for the purposes of improving services or cutting costs. In today's context of functioning of ISARS, its importance has become more. This Unit discusses the basic and general compatibility issues and provides an overall background of problems to be considered in establishing and maintaining an information system.

In resource sharing context of information system, compatibility is a term concerned with the operations of two or more information systems and their ability to work together effectively to exchange records, make use of each other's services and equipment, etc.

The compatibility is capability of being used together without modification or adaptation and is concerned with all work of information systems. It may be noted that compatibility is at the basis of much of the work that has been done to promote use of national or international standards. There is a relationship between compatibility and standardisation. Adoption of same standards is the most obvious, the easiest and the most economical method of ensuring that two or more organisations will be able to achieve compatibility in the particular functions that they will perform together, that is operating as an information system. For example, when considering the exchange of records relating to sources of information, most vital ones are standards relating to the preparation, production and processing of the bibliographic records.

Compatibility can be achieved also, without standardization by the use of some procedures or devices which reconcile or convert one set of records to the format of another system. But, such solutions may introduce further barriers to be overcome, may be more expensive and perhaps less effective.

Therefore, in planning and establishing information systems of any type, it is most desirable to build compatibility into the system from the beginning. The standards for every step of preparation and processing of records should be defined, developed and agreed upon, as well as other standards related to the functions and services that will be part of the operational system need to be considered.

15.2 NEED FOR COMPATIBILITY AMONG ISAR SYSTEMS

In the context of resource sharing, compatibility is more required. The reasons are:

i) it facilitates exchange of bibliographic records, or some means of common access to these records. It is concerned with the recorded description and item of informational context, not the physical item itself.

ii) computer and telecommunication technologies greatly facilitate sharing of resource among systems which, in many cases are impractical in the manual environment such as exchange of large bibliographic files.

iii) in automated environment, the possibilities of cooperation are very much greater for which compatibility becomes very important. Computers can operate efficiently when standards are rigorously followed and compatibility is maintained.

iv) in automated situation absence of compatibility leads to loss of efficiency and results in additional costs in resource sharing.

v) even in the context of Internet, the absence of compatibility leads to inconsistency in the performance of search engines.

In the words of Taylor G. Arlene and Clemson Patrice [2004]:

- there is much duplication of entries within the same set of retrieved hits;
- results are unpredictable;
- results can be quite misleading–the same search can retrieve no hits by one engine and many hits by another;
- search engines do not readily disclose the contents of their databases nor do they provide a description of the criteria used to include a document in their files;
- vocabulary is not controlled, and punctuation and capitalisation rules are not standardized; and
- relationships and relevance often cannot be analysed without actually examining each item—that is, there is not enough information in the 'index entry' to allow one to make educated choices.

These factors are to be considered in relation to the three groups most involved with information and information systems, namely, producers of information, distributors of information and the users of information.

It is with this perspective different areas of compatibility are described and discussed in the text of this Unit.

15.3 SCOPE OF COMPATIBILITY IN ISAR SYSTEMS

Computerised storage and retrieval of information poses several challenges. The choice of a system depends upon its compatibility with the existing data as well as equipment in terms of both hardware and software tools. Accordingly it has direct impact on the decision-making procedure and necessitates bench marking of existing systems and tools before adapting any one system. Further there is fast development in computer technology and it is absolutely desirable that the adapted system is upward compatible with next generation systems and tools. Perhaps the information storage and retrieval systems are most affected by the changes in the Information Technology tools as demonstrated by the profession having the experience from offline CDROM based services to online dialup (DIALOG like services) then onto web based services delivered in real time.

The profession has gone through several forms of storage of information and also several technological advances in packaging and delivering the information through existing communication methods. While the content has also evolved more into interdisciplinary nature than pure subjects; the matter of real concern of the information professionals lies in achieving compatibility firstly at the level of description, medium of storage, processors, operating system and other tools and communication tools.

15.4 AREAS OF COMPATIBILITIES IN ISAR SYSTEMS

Compatibility is promoted in areas of Information Storage and Retrieval Systems when we need some sort of co-ordination between several of them. Compatibility is to be achieved between ISARS working in different environment and conditions. For example, ISARS in USA and India are a case in point. One way is to directly convert an output of one system (A) into a form suitable for use as information to a second system (B) as for example when a computer program is written to convert the record format of A into a format that can be operated in B's computer. Another way of compatibility target among organizations is from the angle of the establishment of some type of mediating device that acts

as a kind of switching mechanism, converting an output of any participating centre into a form usable by any other participating centre. Thus, a **Common Communication Format**, can transform a processing format used by A into one usable by B (and C, D, and so on) and vice versa. An intermediate lexicon can perform the same functions between vocabulary controls of different organisations. Computer and communications have to be adopted to meet the demands of information service. For this purpose, the automation of information services makes the achievement of compatibility more important than ever before.

Information systems are normally involved in manipulating records representing information sources such as books, periodicals, periodicals articles, and other carriers of information (such as non-print or paper materials). It is in these records and the way they are handled, rather than in the physical items represented or the handling of the physical items, that compatibility is so important. This applies to records in machine-readable form.

Some of the areas of compatibility can be:

i) **Systems and Components**

The present corporate, scientific, research and academic world is knowledge hungry and information systems often contain several terabytes of data. Typically these systems do not operate offline but must make available the terabytes of information online with very efficient retrieval engines and user-friendly interfaces.

There must be both downward and upward compatibility as far as the medium as well as the data storage formats are concerned. Prescription to standards companies dealing with tools and systems is one built-in mechanism to ensure compatibility. Most leading companies ensure the compatibility while the same may not be possible for tools from smaller and more little known or super specialised companies. When selecting a storage system the information professionals are well advised of factors to be considered such as, availability, reliability, interoperability, scalability, costs, optimal performance, etc.

ii) **Record Format**

The compatibility in bibliographic record format has been a live issue during the last 30 years. The development of ISO 2709, UNISIST Reference manual, UNIMARC, MARC21 and Common Communication Format are some of the efforts in addressing important compatibility issues for international standards and representation in this field.

iii) **Data Formats**

Data format by definition is a specified arrangement for storage or retrieval. Several applications produce and interpret data only in proprietary formats. This implies that to read files produced by that application it must be obtained and installed on the system. As against this, formats such as plain text are compatible across applications and directly processible by systems. Popular examples are the computer programs. However, there are several types in binary, plain, image, video and audio formats.

Text formats can be of following types:

Simple Text, such as, texts using ASCII, ISCII, and UNICODE character codes.

Structured Text Format, such as, Standard Generalized Markup Language (SGML), Hypertext Markup Language (HTML), Extensible Markup Language (XML), etc.

Page description languages, such as, postscript, portable document format (PDF), Tex, etc.

Image formats can be of following types:

Graphic Image Files (GIF), Joint Photographic Experts Group (JPEG/ JPG/ JIF), Targa File Format (TGA), Tagged Image File Format (TIFF), Bitmap Image (BMP), etc.

Video file formats can be of following types:

Audio/Video Interleaved (AVI), Movie (MOV), Moving Picture Experts Group (MPEG/ MP3), etc.

Audio file formats can be of following types:

Audio Interchange File Format (AIF), Moving Picture Experts Group (MPEG/ MP3), Musical Instrument Digital Interface (MIDI), VOC Format, WAV Format, etc.

iv) **Media Formats**

There has been a continuous change in the handling of information due to the influence of changing technologies and techniques. Information in different forms was introduced such as microfilms and negative films, cine films and videotapes, discs and sound tapes, microcomputers and optical storage systems.

There are a number of storage media formats, such as, hard disk, floppy disk, CD-ROM, DVD, zip drive, cartage tape, pen drive, etc. that hold data for archival purpose, data backup and online systems. The storage media formats should be compatible across the computer systems, operating systems and applications, so that same data can be reusable. There are communication media formats, where data is transfered across communication media. Here sender's messages should be understood by the receiver's system. For data security reasons, sometimes encrypted data is transmitted from one system to another system. Receiver should have a mechanism to decrypt the data after authentication.

v) **Retrieval Formats**

The storage and retrieval formats generally are different. It is desirable in an IR system to provide for versatile retrieval formats as desired by the users. Generally bibliographic database management system provide for data retrieval in standard format such as AACR2 format.

Query formulation formats play an important role in helping users frame meaningful queries so that the noise factor in retrieval can be reduced. Query formulation must be aided with adequate vocabulary control devices and navigational facilities to help user formulate or re-formulate queries. Intuitive but non-intrusive assistance should be built into the querying system.

The other factor is that of display. An efficient retrieval should result in a meaningful display of the retrieved set. The display features of many search engines leave much to the imagination of end users or sometimes give unnecessary details.

vi) **Interface Compatibility**

Various compatibility issues in ISARS arise in an effort to make the search systems user friendly. The compatibility issues are :

a) can the user terminal communicate with the computer?

b) can all the databases be searched in the same way?

c) can a search strategy be transferred from one database to another?

d) are all the results obtained from ISAR's understandable to users?

e) can outputs be merged and duplications eliminated?

f) can the search outputs be stored in machine-readable form by the user?

In case where more than one computer or bibliographic/textual databases are studied, the user terminal had to communicate with each one of them. This raises the interface for compatibility of information exchange. To be user friendly search, we look to a single command language. This can help a common search through several databases.

15.5 PRINCIPAL ISSUES OF COMPATIBILITY IN ISAR SYSTEMS

The main objectives, structure and function of ISAR systems centre round the following aspects and compatibility is looked for in these aspects:

a) Generation of facts and ideas in compatibility with information requests from the searchers of the system;

b) Provide factual quantitative and qualitative data of information to meet needs for users;

c) Processes of creation, repackaging, or particular kinds of information channels;

d) Processes of creation of particular kinds of documents as communication channels as desired by searchers;

e) Processes of the flow of information carriers to the immediate communication needs;

f) Information banks, their parameters and structure, technological processes of information storage, retrieval, and dissemination;

g) Human element within the systems of originating, and using information the roles of generators, receptors, and mediators of information communication;

h) Information system elements, structural relations and behaviour;

i) Organisation, managing and controlling of information systems and networks and information between systems;

j) Technical compatibility of communication processes and operations;

k) Information languages, natural and artificial in their syntactic, semantic and pragmatic aspects.

Compatibility solutions are being searched for ISARS through multi-disciplinary approaches through the methodologies available in the fields of cybernetics, logic, semantics, linguistics, sociology, psychology, etc. These are leading to variety of cognitive paradigms for compatibility for ISAR systems.

15.5.1 Interoperability

Interoperability refers to the ability of information systems to operate in conjunction with each other encompassing communication protocols, hardware, software, application, and data compatibility layers. It also means the ability of different types of computers, networks, operating systems, and applications to work together effectively, without prior communication, in order to exchange information in a useful and meaningful manner. There are three aspects of interoperability: semantic, structural and syntactical. Interoperability is the ability of one manufacturer's computer equipment to operate alongside, communicate with, and exchange information with another vendor's dissimilar computer equipment. Service

interoperability is the possibility for different services to operate on different networks. The technical specifications at the interconnection interface determine, in part, service interoperability between different operators. Interoperability in information systems refers to the ability of two or more systems to exchange information and to mutually use the information that has been exchanged. For example, if the library OPAC records are in the ISO-2709 format and use the same bibliographic data formats such as MARC21 for bibliographic description then the two are truly interoperable and data can be freely exchanged between the two.

15.5.2 Scalability

Information explosion is indeed a clichéd term in the information service profession as 100 Terabyte online libraries are most common and being added everyday. Several databases are being replicated for disaster recovery and network contingencies. Hence a system designed for such information systems must be scalable to support large databases and still work at an optimal speed of retrieval.

Self Check Exercises

1) Identify the principle issues of compatibility of ISAR system.

2) What is Interoperability?

Note: i) Write your answers in the space given below.
ii) Check your answers with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

15.6 COMPATIBILITY OF ONLINE IR SYSTEMS

Web has remarkably influenced transactions on the Net and made it truly interoperable. In the sense that when we are browsing the net and click on links and navigate between sites we seldom bother to know what operating system the distant machine may have because in the web environment it really does not make a difference. Hence, even if the library automation software or digital library software is configured on a UNIX platform the end users operations can be through web browsers so that users can be comfortable in familiar environment. Online IR systems are adhering to international and world standards to achieve true interoperability. MARC21 standard is followed for library catalogues. Most tools developed for data exchange between libraries are also geared up to the MARC21 standard.

15.6.1 Z39.50 Standard and Protocol

Z39.50 standard and protocol enables exchange of bibliographic record between libraries even though they may have different description standards. The standard is designed to facilitate interoperability between computer systems whether on Internet or on Intranet. The protocol does not deal with aspects of interaction between user and the origin or target machine. It only deals with interaction between origin and target machine. The essential functions are searching and retrieving information from databases available on multiple hosts. The protocol basically specifies data structures and interchange rules that allow a client machine (origin) to search databases on a server machine (target) and retrieve records.

In a typical implementation of Z39.50, the origin and the target should somehow translate their messages into a common language. Both the origin and target should be Z39.50 compliant. The origin's search query should be mapped as Z39.50 query. On the target side the Z39.50 query should be mapped as its database query and the results are presented.

The Z39.50 standard does not broadcast searches to multiple servers, but a client can open Z39.50 sessions with multiple servers either sequentially or simultaneously. However, manipulating multiple results from different targets, removing duplicates and presenting retrieved records in a uniform fashion to the end-user are not covered by the protocol [http://lcweb.loc.gov/z3950/].

On the target side, the Z39.50 gateway resides on the web server (like Apache or IIS). In addition, browser-based implementations using either Java or Active X applets reside on the target and are to be downloaded to the user's machine.

As no two databases are expected to be alike with regard to the structure (data elements) and searchable fields, it is required to develop a common abstract model of the target databases. The model should contain the abstract data structure (schema) having the data elements like author, title etc and also the searchable elements as all data elements need not be indexed.

Although Z39.50 is not a database indexing standard, Z39.50 profiles developed for specific communities require a commonly agreed upon database indexing standard. These profiles normally include a minimum set of access points and they should be supported by the database indexes to ensure interoperability between target systems.

The Z39.50 protocol also helps in:

- identifying the characteristics of the server data base; and
- locating the databases distributed across the Internet.

15.6.2 Interoperability of Online Repositories

In the area of Internet resources, Dublin Core (DCMES) is followed as a world standard for description of web resources. The Dublin Core metadata standard is a simple, yet effective element set for describing a wide range of networked resources. The Dublin Core standard comprises fifteen elements, the semantics of which have been established through consensus by an international, cross-disciplinary group of professionals from librarianship, computer science, text encoding, the museum community, and other related fields of scholarship. The Dubline Core has been discussed in Block 2 of this course.

To enable meaningful retrieval by search engines it is necessary to semantically enrich the web resources. The OAI-PMH (Open Access Initiative – Protocol for Metadata Harvesting) advocates a simple set of rules for exposing metadata of digital repositories to search engines so that meaningful retrieval is possible. A Static Repository is an XML file that is made accessible at a persistent HTTP URL. The XML file contains metadata records and repository information. A Static Repository becomes accessible via OAI-PMH through the intermediation of a Static Repository Gateway.

The Open Archives Initiative (OAI) has focused on developing a framework where metadata can be harvested from multiple digital library collections or 'repositories'– and can then be loaded into a centralised service. The result is a database of metadata records that has been built from multiple, related repositories. Then, as users search the new database and select metadata records that interest them, the service provides links that allow them to view the

corresponding object on the original repository [Breeding, 2002]. In general terms, this relates to enabling general search engines to search OAI complaint database and retrieve information from structured databases or library collections in addition to searching plain HTML pages.

Compatibility is a concept. No specific method can be prescribed for achieving compatibility. In whatever way two or more systems can be developed by which it is possible to share and interchange the resources decide the methodology.

15.7 APPROACHES TOWARDS COMPATIBILITY IN ISARS

Some of the methodologies that can be seen are:

i) theory of scientific information as an instrument for organisation, management and control of the sources of human knowledge search of compatibility through epistemology function;

ii) structural compatibility of information of a new general method of science studying the diffusion of information from the point of view of generation, circulation and operation of information associated with human development phenomena;

iii) information system analysis and synthesis;

iv) content analysis and transfer of semantic content (not only measuring the ratio and the amount of transferred information);

v) methods of bibliographic organisation and control of human knowledge;

vi) specific techniques and methods for selection of relevant information;

vii) process of reduction and surrogation without loss of intellectual content of information.

15.7.1 Multi-Task Approaches

In compatibility process of ISARS, a multi-task approach can be delineated as follows:

i) All specific fields of information and communication activities, namely the methods, techniques, and organisation (including institutional, regional and national cooperation) of generation (handling and processing), storage, search and dissemination of information contained in the documents;

ii) Problems of effectiveness, scientific character standardization, and legal aspects of information and documentation;

iii) Information and documentation promotion activities;

iv) The history of information and the scientific compacting processes;

v) The essence, structure, and dynamic regularities and also interrelationships between different variety of sources of documented information; and

vi) The essence, structure, and dynamic regularities of information seekers' needs in information contained in documents and specific methods for study and classification of these structures.

Automation is considered as a solution to handle the vast information storage and retrieval process. But introduction of even the most efficient computer together with the necessary communication and reprographic automation and consequently creation of man-machine systems in social information may raise many new intellectual, socio-psychological, neuro-physiological and other problems, that can be solved by cooperation of researchers and practitioners from different fields of knowledge.

Self Check Exercise

3) Enumerate different tasks involved in compatibility process of ISARs.

Note: i) Write your answer in the space given below.
ii) Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

15.7.2 Interactive Planning

Interactive planning has several principles which may facilitate compatibility process between ISAR systems. The process of planning is more important than the actual plan produced. The first principle is the participation process. The idea is that all those who are likely to be affected by the systems should be in the systems planning and development. It leads to coordinated activity leading to organizational integration and motivation. These lead on to the compatibility.

The second principle is that of continuity. The participants, particularly of ISAR systems need change over time and this will necessitate corresponding changes in plans. Surprise events may occur, chaos will emerge. The system may not work as planned or other changes in the organization environment may change the situation in which it finds itself. Thus, principle of continuity should constantly search for compatibility, based on a continuous feedback of changes.

The final principle is the holistic principle. The compatibility process should emphasize simultaneity and inter-dependence for parts and levels of the system. This leads to a principle of coordination which states that units at the same level should plan together and at the same time because it is the interactions between units rather than interdependent actions that give rise to most difficulties; and the other principle is the "principle of integration" which insists that units at different levels plan simultaneously and together, because decisions taken at that level usually have effects on the other systems.

The interactive planning methodology operates in five phases namely:

i) Formulating of the Mess

ii) Ends Planning

iii) Means Planning

iv) Resource Planning

v) Implementation and Control

Formulation of the Mess

It consists of identifying and analysing the problems of opportunities, prospects and organisation for achieving the objective. According to Ackoff, it has three types of projections.

a) System analysis: giving detailed picture of the organisation and how it works, whom it affects and how, and its relationship with its environment;

b) An obstruction analysis: setting out any obstacles to corporate network system developments; and

c) Preparation of reference projections: which extrapolate on the system's present performance in order to predict future performance. If nothing is done, trends in the environment continue as they are.

Synthesising the results of these three types of study yields a reference scenario which is formulation of the mess currently in.

Ends Planning

It concerns with specifying the ends to be achieved in terms of the ideals, objectives and goals. The process begins with 'idealised design' which is both unique and the most essential feature of Ackoff's approach. An idealized design is prepared by going through the following steps:

a) *Selecting a mission* — which is general purpose statement incorporating organisation's responsibilities to its environment and propounding a vision of what the organisation could be like which generates commitment;

b) *Specifying desired properties of the design* — a comprehensive list of the desired properties which users agree to be built into the system; and

c) *Designing the system* — setting out as to how all the specified properties of the idealised design can be achieved.

An idea-seeking system requires a particular kind of organisational design capable of rapid and effective learning and adaptation. Ackoff provides an outline for such a 'responsive decision system'. This contains five essential functions:

a) Identification and formulation of problems (Threat/Opportunities);

b) Decision-making – determining what to do abut the threats and opportunities;

c) Implementation – doing it;

d) Control and monitoring performance and modifying actions to prevent repetition of mistakes; and

e) Acquisition or generation and distribution of the information necessary to carry out the other functions.

Those organisations willing to undertake idealized design should, according to Ackoff, reap considerable benefits. In particular, the process is said to control compatibilities by means such as :

a) Facilitate the participation of all users in the planning process;

b) Allow incorporation of aesthetic values of the information users in the planning of ISARS;

c) Generate consumers among those who participate;

d) Release large amount of suppressed creativity and harness it to the advantage of individual and organisational development;

e) Expand participant's concept of feasibility revealing that the biggest obstruction to the future we most desire ourselves; and

f) Ease of implementation, since people are more inclined to implement plans in which they have a say.

Means Planning

It is the third phase after the reference scenario (phase I) and idealised design (phase II). During phase III, means planning, policies and proposals are generated and examined with a view to decide whether they are capable of helping to fill the gap between the desired future and the way future looks like at the beginning.

Creativity is needed to discover ways of bringing the organisation towards the desirable future, invented by its users. Alternative means to reach the specified ends must be carefully evaluated and selection made. Compatibility issues are to be looked with greater care and vision.

Resource Planning

During the stage of planning, Ackoff recommends that four types of resource should be taken into account :

a) Inputs — materials, supplies, energy and service

b) Facilities and equipment — capital investments

c) Personnel

d) Money

These are to be interacted and analysed with the means adopted in the earlier phase. Compatibility factors for ISAR get into a critical stage in this phase.

Implementation and Control

This final phase of interactive planning concerns itself with the implementation of decisions made in the earlier stages. Who is to do what, when, where and how implementation is continued, should be monitored and with feedback process.

Self Check Exercise

4) Enumerate the resources that are to be considered for compatibility during the Resource Planning Phase.

Note: i) Write your answer in the space given below.
ii) Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

15.8 QUALITY CONTROL AND COMPATIBILITY

The overall compatibility issues in ISAR systems ultimately centre-round quality concepts. Here we can quote Manfred Kochen [1976] : "An information system may be used but not be useful; it may also be useful and not used. It may even be neither useful nor used. It is ideal if it is both used and useful".

Compatibility keeps close with standards. Standards base themselves on measures of quality and quantity. The following quotations from Daniel [1993] indicate the studies on these aspects. To assess feasibility and desirability, quality control issues as applied to the organisation as a whole, will be reviewed. Lessons gleaned from this material will be applied first to qualifying issues dealing with the selection and indexing of documents for inclusion in a database that is, what determines quality in the development of a product or service on the input side. The second aspect of quality control of documents focuses on throughout, that is what determines quality in maintaining currency, and relevancy within the collection of databases.

A third assessment of quality relates to the selection of material from a collection or database in response to a user's need, that is, what determines quality on the output side in the delivery of a product or service. We can further quote Daniel on the quality control to emphasise its correlation to compatibility.

The adoption of a TQM orientation to quality demands emphasis on customer defined standards, where customer is interpreted to mean whoever is the receiver of a product or service, whether internal or external, quality must be judged by the customer. The quality management approach focuses on finding and reducing causes of variation that occur in the systems used to produce the results. We must form partnerships with all other participants in the chain of database producer to end-consumer and plan a total system commitment to continuous quality improvement.

15.9 SUMMARY

Compatibility of Information Storage and Retrieval Systems (ISARS) refers to the property that facilitates two or more organisations to communicate and exchange data. Compatibility between ISAR systems are judged on the basis of the structure and functions performed by the system.

Thus, during the intermediary interaction with the searcher one builds a typology of user model. The interpersonal communication of intermediary with potential users is very important. Usually, the educational level of the users, the type of question, the way the question is phrased, the purpose of the search, and the expected result of search – all play a major role in information service. The intermediary, in essence, builds a user profile and specifies the task requirement process. The user-modelling has played an important role in tuning the database they are having in action. An intelligent interface for document retrieval systems should exhibit for interaction with users and the reaction obtained for improvement of database searching. Frame-based representation does help in defining user modeling and provides pre-search and post search productivity. We can also build models of typical or untypical users, develop stereo types wherever possible and provide interconnection to their information needs and a dialogue with colleagues.

Compatibility of ISAR systems can be achieved either by coordination among several systems or through some mediating device. The mediating device acts as a switching mechanism which converts data output of one centre into a form usable by any other participating centre. The Common Communication Format (CCF) is one such example of a mediating device.

To achieve compatibility of ISAR systems it is essential that they are based on certain set standards. Standardisation of various structure and functions of ISAR system can help in achieving quality. The utility of ISAR systems depend on the compatibility and quality aspects, which can only be achieved through standardisation. ISO 2709, UNISIST Reference Manual, UNIMARC and Common Communication Format, Z39.50, Dublin Core are some of the examples of international standards that focuses on compatibility issues.

Compatibility is essential for information exchange. With increasing storage of information in machine-readable formats and dissemination of information on the multimedia format, compatibility is becoming an important issue. Compatibility aspect needs to be tackled right from the planning stage of the information systems and interactivity between different system in essential to make communication and exchange of ISAR systems possible.

15.10 ANSWERS TO SELF CHECK EXERCISES

1) The compatibility of ISAR systems should be based on :

   i) users need for information

   ii) processing of information and documents according to the user needs

   iii) flow of information carries to immediate information needs

   iv) storage, retrieval and dissemination aspects of information banks

   v) human element as generator, mediator, and user of information

   vi) information system elements

   vii) management of information system

   viii) technical compatibility of communication process and operation

   ix) information languages used.

2) Interoperability refers to the ability of information systems to operate in conjunction with each other encompassing communication protocols, hardware software, application, and data compatibility layers. It also means the ability of different types of computers, networks, operating systems, and applications to work together effectively, without prior communication, in order to exchange information in a useful and meaningful manner. Interoperability in information systems refers to the ability of two or more systems to exchange information and to mutually use the information that has been exchanged. For example, if the library OPAC records are in the ISO2709 format and use the same bibliographic data formats such as MARC21 for bibliographic description then the two are truly interoperable and data can be freely exchanged between the two.

3) The different tasks involved in the compatibility of ISAR systems are :

   i) Methods, techniques and organisation involved in generation, storage, search and dissemination.

   ii) Dealing with problems like standardisation, legal aspects, effectiveness etc.

   iii) Promotional activities

   iv) Scientific compacting processes

   v) Compatibility between variety of sources of information

   vi) Matching of information needs of the users with the information contained in the documents.

4) The four types of resources needs to be taken into consideration while planning for ISAR systems. They are :

   i) Inputs — materials, supplies, energy and service

   ii) Facilities and equipment — capital investment

   iii) Personnel

   iv) Money

15.11 KEYWORDS

Common Communication Format (CCF) : It is bibliographic exchange format useful for libraries and information centres developed by UNISIST.

Compatibility of Information System : Adaptability of Information Systems in various environments.

Dublin Core : A set of 17 elements prescribed for description of web resources

Interoperability : Interoperability refers to the ability of information systems to operate in conjunction with each other encompassing communication protocols, hardware software, application, and data compatibility layers.

TQM : Total Quality Management is a modern technique of quality management and quality assurance that takes care of performance of whole system and its components or subsystems.

UNIMARC : Universal machine-readable catalogue — a standard development under the auspices of a working group set up by IFLA.

15.12 REFERENCES AND FURTHER READING

Arlene, T.G. and Patrice, C. *Access to networked documents: catalogs? search engines? both?* Position Paper, OCLC Internet Cataloging Project Colloquium. <http://www.oclc.org/oclc/man/colloq/taylor.htm> (browsed on 01/07/2004).

Breeding, Marshall. (2002). Understanding the protocol for metadata harvesting of the Open Archives Initiative. *Computers in Libraries*, 22(8), 24.

Brickley, Dan. *Z39.50 resources. <http://www.ilrt.bris.ac.uk/discovery/z3950/resources>.*

Danial, Evelyn H. (1993). Quality control of documents. *Library Trends*, 41(4), 646-64.

Dublin Core Metadata Element Set. <http://www.dublincore.org/documents/dces>.

Frater, Harald and Paulissen, Dirk. (1994). *Multimedia mania*. Grand Rapids: Abacus.

Gopinath M.A. (1999). *Compatibility of ISAR* systems. In: MLIS-03 Course Material, Block-3, Unit 11. New Delhi: IGNOU.

Information Retrieval (Z39.50): ANSI/NISO Z39.50-1995 application service definition and protocol specification. <http://lcweb.loc.gov/z3950/agency/1995doce.html>.

Jeff, Saake. *DVD-RAM for the new millennium.* <http://www.panasonic.com/industrial/computer/storage/dvdram/new/article/art64.htm> (browsed on 11th July 2004).

Kardorf, B. (1998). SGML and PDF: why we need both. *Journal of Electronic Publishing*, 3(4), 14p. <http://www.press.umich.edu/jep/03-04/kardorf.html> (browsed on 23rd July 2004).

Kochen, M. and Donohue, J.C. (1976) (Eds.). *Information for the community*. Chicago: ILALP.

Rosch, Winn L. (1995). *Multimedia Bible*. Indianapolis: Sams Publishing.

Woodward, H. [et al]. (1977). Electronic journals: myths and realities. *OCLC System and Services,* 13(4), 144-151.

UNIT 16 INTELLIGENT INFORMATION RETRIEVAL SYSTEMS

Structure

16.0 OBJECTIVES

Development of several IR systems using artificial intelligence techniques have been attempted. In this Unit, we shall try to apprise you with their details. The Unit discusses a few Expert Systems developed for information services. Further the lesson aims to describe the components of expert systems with emphasis on knowledge representation and knowledge engineering and also the systems developed using pattern recognition techniques.

After reading this Unit you will be able to:

- know the meaning of 'Artificial Intelligence';
- gather a view on the state of the art of intelligent IR system;
- appreciate the applications of intelligent systems in information processing and organisation;
- understand the features of expert systems, decision support systems and AI-based decision support systems; and
- understand the features of Pattern Recognition and its applications in IR system.

16.1 INTRODUCTION

The main aim in Artificial Intelligence (AI) research is to develop computer programs that behave like humans. Such systems are expected to think, solve problems and make decisions in an intelligent manner. The term *intelligence* covers many cognitive skills, including the ability to solve problems, learn, and understand language – AI addresses all of these. But most progress to date in AI has been made in the area of problem solving –concepts and methods for building programs that *reason* about problems rather than calculate a solution.

According to Rich, "Artificial Intelligence is the study of how to make computers do things at which, at the moment, people are better" [Rich, 1990]. Patterson [1992] defines AI, "as a branch of computer science concerned with the study and creation of computer systems that exhibit some form of intelligence, systems which learn new concepts and tasks, systems that can reason and draw useful conclusions about the world around us, systems that can understand a natural language or perceive and comprehend a visual scene, and systems that perform other types of feats that require human types of intelligence".

The advantages of a system that has human intelligence are many. Intelligent systems can be used wherever human intelligence is required, like in business, engineering, manufacturing, medicine, etc. In addition, such a study has intrinsic advantages, as it demands understanding and meticulous analysis of the human thinking process, so as to develop computational models. The following reasons, given by Winston [1984], justify the study of AI as means of understanding human intelligence.

1) The use of computers demands a clear statement of the problem and clear strategy for solutions and this requires a clear understanding about the human thinking process.
2) Computer models force precision. Implementing a theory uncovers conceptual mistakes and oversights which ordinarily escape the notice of even the most meticulous researcher.
3) Computer implementations qualify task requirements. Once a program performs a task, upper bound statements can be made about how much information processing the task requires.
4) It is usually simple to deprive a computer program of some piece of knowledge in order to test how important that information really is.

Self Check Exercise

1) What are the factors that lead to understanding 'Intelligence'?

Note: a) Write your answers in the space given below.
b) Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

16.1.1 Branches of Artificial Intelligence (AI)

In one of the first attempts, AI scientists tried to simulate the complicated process of thinking by applying general methods for solving broad classes of problems. These programs are called the 'General Purpose Problem Solving (GPS)' programs, developed in the 1960s. These programs were expected to handle more classes of problems based on some general rules. But this strategy could not produce major breakthroughs, as the programs could not reach conclusions, by applying general rules, in many cases. In the words of Newell [1969], "General purpose problem solving programs are too weak to solve most complex problems". The limitations of very general strategies adopted gave way to efforts in narrowly defined application areas.

Accordingly, attempts in the early seventies were to build more specialized

programs that used general methods or techniques. The efforts were to build comprehensive knowledge representation theories and associated general-purpose systems [Borrow and Winoguard, 1977; Minsky, 1975]. But these efforts also faced the same limitations as earlier. It was realized that 'knowledge' as a target of study is too broad and diverse and efforts to solve knowledge-based problems in general are premature.

In the late 1970s came the breakthrough realization that the performance of intelligent programs depends upon the knowledge it has – "the power of an expert system derives from the knowledge it possess, not from particular formalisms and inference schemes it employs"[Feigenbaum, 1998]. An intelligent program depends upon lots of high quality '*specific*' knowledge about some problem area [Hayes-Roth, Waterman and Lenat, 1983]. This led to the development of special purpose computer programs. These programs were *'domain specific'*, that is, they could be experts in some narrow problem area. These programs are called 'Expert Systems' and the process of building expert systems is called 'Knowledge Engineering'. The objective is to build effective programs that solve problems based on expert knowledge furnished. The knowledge may be simple facts, procedural information and strategies.

Artificial Intelligence is a subject by itself, though it can be considered as a branch of computer science. It is a highly interdisciplinary area influenced by many other areas like cybernetics, linguistics, epistemology, logic, information theory and psychology. The main areas under artificial intelligence are:

- Natural Language Processing (NLP)
- Expert Systems (ES)
- Pattern Recognition
- Robotics

The NLP-based system is thought to be of relevance to textual information retrieval. Details about NLP-based IR systems have been discussed in Unit 4.

16.2 EXPERT SYSTEMS

Expert systems are computer programs that are derived from a branch of computer science research called *Artificial Intelligence* (AI). AI's scientific goal is to understand intelligence by building computer programs that exhibit intelligent behavior. It is concerned with the concepts and methods of symbolic inference, or reasoning, by a computer, and how the knowledge is used to make those inferences will be represented inside the machine. In the following paragraphs let us study the views of experts in understanding the concept of expert systems.

Specialised computer programs, modeled on the way human experts tackle problems and arrive at solutions are referred to as 'Expert Systems'. Such systems rely upon a store of specialized knowledge for solving problems and hence also referred to as Knowledge Based Computer Systems (KBCS) or Knowledge Based systems (KBS).

Expert systems are sophisticated computer programs that manipulate knowledge to solve problems efficiently and effectively in a narrow problem area. Knowledge based systems enhance the value of expert knowledge by making it readily and widely accessible. Like human experts, these systems use symbolic logic and heuristics to find solutions. They are also capable of learning from experience through inferencing mechanism. According to Liebowitz 'the role of experts systems is to better understand how humans think, reason and learn' [Leibowitz, 1990].

AI programs that achieve expert-level competence in solving problems in task areas through the knowledge about specific tasks are called *knowledge-based* or *expert systems*. Often, the term expert systems is reserved for programs whose knowledge base contains the knowledge used by human experts, in contrast to knowledge gathered from textbooks or non-experts. More often than not, the two terms, expert systems (ES) and knowledge-based systems (KBS) are used synonymously. Taken together, they represent the most widespread type of AI application. The area of human intellectual endeavor to be captured in an expert system is called the *task domain*. *Task* refers to some goal-oriented, problem-solving activity. *Domain* refers to the area within which the task is being performed. Typical tasks are diagnosis, planning, scheduling, configuration and design.

Building an expert system is known as *knowledge engineering* and its practitioners are called *knowledge engineers*. The knowledge engineer must make sure that the computer has all the knowledge needed to solve a problem. The knowledge engineer must choose one or more forms in which to represent the required knowledge as symbol patterns in the memory of the computer – that is, he (or she) must choose a *knowledge representation*. He must also ensure that the computer can use the knowledge efficiently by selecting from a handful of *reasoning methods*.

16.2.1 Expert Systems for Information Processing and Retrieval

A few expert systems in the field of information processing, information analysis and retrieval are discussed below.

COMIT: COMIT is an early AI language developed at MIT (Massachusetts Institute of Technology). The program is designed for information retrieval. It matches the contents of a highly structured query to highly structured descriptions of biographical sources [Weil, 1968]. A categorization scheme is used as a basis for inferencing and the system follows a pattern matching procedure. Reference sources are described in functional rather than bibliographic terms. The hits or successful matches are ranked according to the degree of certainty that they could answer a query. A declarative knowledge representation scheme is employed for COMIT.

REFSEARCH: REFSEARCH was developed by Joseph Meredith and team, and attempted representation of entire universe of reference works. This program is developed as an instructional tool for teaching librarians the basic principles of reference work. The system aims to study the "principles that would apply to the collection as a whole, to the sum of the data contained in the collection, and to networks of paths leading to the data" [Meredith, 1971]. REFSEARCH describes the various reference sources in terms of the functions they performed. The focus is to describe the reference works as specific types of tools likely to resolve types of information problems. The schemata for knowledge representation in REFSEARCH is quite similar to the methods adopted in present expert systems.

RESEDA: RESEDA developed by Zarri, is a system based on restructuring the knowledge encoded in printed texts. The Reseda system is based on the principles of linguistics. The purpose of RESEDA is recording information found in standard printed works on medieval French biography and history. The goal of Reseda is to use the knowledge base derived from the texts to answer natural language queries about the domain [Zarri, 1985].

CANSEARCH: Politt [1986, 1987] has developed Cansearch, an expert system for access to cancer therapy literature in the Medline database. Cansearch is a hybrid system using rules and frames for knowledge representation. Controlled vocabulary terms and hierarchical relationships between concepts expressed in

the Medical Subject Headings are used to guide the search process. Queries formulated are based on knowledge of typical searches, concepts within the domain and the Medline query language.

IR-NLI (Information Retrieval – Natural Language Interface): IR-NLI translates the user's initial query into a formal problem statement. It then models the intermediary's behaviour by consulting a knowledge base of expert intermediary knowledge. The knowledge base contains rules relating to tactics, strategies, and approaches used by searchers. Domain knowledge fed is terminological knowledge about the subject domain of the target database. A formalizer module then generates formal, syntactically correct search strategies [Brajnik and Tasso, 1986].

16.2.2 Components of Expert Systems

In general, expert systems are composed of basic components such as:

a) a user interface: to facilitate user interaction

b) a knowledge base: the facts or knowledge based upon which the ES makes decisions

c) an inference mechanism: the reasoning engines built according to heuristics reasoning or facts.

The *knowledge base* of expert systems contains both factual and heuristic knowledge:

Factual knowledge is that knowledge of the task domain that is widely shared, typically found in textbooks or journals, and commonly agreed upon by those knowledgeable in the particular field.

Heuristic knowledge is the less rigorous, more experimental, more judgmental knowledge of performance. In contrast to factual knowledge, heuristic knowledge is largely based upon past experiences and knowledge gained. For instance the knowledge that, if the same pattern of events take place then the same conclusion could be expected. It is the knowledge built from 'thumb rules'.

Self Check Exercises

2) What are the components of Expert Systems?

3) What are the two types of knowledge contained in Expert Systems?

Note: i) Write your answers in the space given below.
ii) Check your answers with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

.....

16.2.3 Knowledge Representation

The knowledge of a given situation or the context plays a crucial role in determining how the system interprets a particular sentence. The different kinds or levels of knowledge required for natural language understanding are [Brajnik and Tasso, 1986]:

Phonetic and Phonological Knowledge: This concerns how words are sounded. This is dealt within voice/speech recognition system.

Morphological Knowledge: Concerns with how words are constructed out of more basic meaning units called morphemes. That is what is referred to as 'root' of the word. For example: The words 'eagerly' or 'playful' are constructed by suffixed letters to the root as shown below:

root: eager suffix: ly = eagerly

root: play suffix: ful = playful

Syntactic Knowledge: Concerns with how words can be put together to form valid sentences of a language. Generally this deals with grammatical part of languages. Grammar sets the rules for framing valid sentences for a language.

Semantic Knowledge: Concerns with the meaning of the words and what senses a string of words in a sentence make.

Pragmatic knowledge: It is concerned with how sentences are used in different contexts to mean differently. This is to add additional information though not expressed in words but inherent in the context.

World Knowledge: It includes the general knowledge about the structure of the world that language users must have in order to infer from a sentence by having prior knowledge about the situation or concerned concepts.

Knowledge representation formalizes and organizes the knowledge. One widely used representation is the *production rule*, or simply *rule*. A rule consists of an IF part and a THEN part (also called a *condition* and an *action*). The IF part lists a set of conditions in some logical combination. The piece of knowledge represented by the production rule is relevant to the line of reasoning being developed if the IF part of the rule is satisfied; consequently, the THEN part can be concluded, or its problem-solving action taken. Expert systems whose knowledge is represented in rule form are called *rule-based systems*.

Another widely used representation, called the *unit* (also known as *frame*, *schema*, or *list structure*) is based upon a more passive view of knowledge. The unit is an assemblage of associated symbolic knowledge about an entity to be represented. Typically, a unit consists of a list of properties of the entity and associated values for those properties.

Since every task domain consists of many entities that stand in various relations, the properties can also be used to specify relations, and the values of these properties are the names of other units that are linked according to the relations. One unit can also represent knowledge that is a 'special case' of another unit, or some units can be 'parts of' another unit.

The *problem-solving model,* or *paradigm*, organises and controls the steps taken to solve the problem. One common but powerful paradigm involves chaining of IF-THEN rules to form a line of reasoning. If the chaining starts from a set of conditions and moves toward some conclusion, the method is called *forward chaining*. If the conclusion is known (for example, a goal to be achieved) but the path to that conclusion is not known, then reasoning backwards is called for, and the method is *backward chaining*. These problem-solving methods are built into program modules called *inference engines* or *inference procedures* that manipulate and use knowledge in the knowledge base to form a line of reasoning.

The *knowledge base* that an expert uses is what he learned at school, from colleagues, and from years of experience. Presumably the more experience he

has, the larger his store of knowledge. Knowledge allows him to interpret the information in his databases to advantage in diagnosis, design, and analysis.

Because an expert system uses uncertain or heuristic knowledge (as we humans do) its credibility is often in question (as is the case with humans). When an answer to a problem is questionable, we tend to want to know the rationale. If the rationale seems plausible, we tend to believe the answer [Feigenbaum, McCorduck and Nii, 1998].

The most important ingredient in any expert system is knowledge. The power of expert systems resides in the specific, high-quality knowledge they contain about task domains. Because of the importance of knowledge in expert systems and because the current knowledge acquisition method is slow and tedious, much of the future of expert systems depends on breaking the knowledge acquisition bottleneck and in codifying and representing a large knowledge infrastructure.

16.2.4 Knowledge Engineering

Knowledge Engineering is the art of designing and building expert systems, and knowledge engineers are its practitioners. There are two ways to build an expert system. They can be built from scratch, or built using a piece of development software known as a *'tool'* or a 'shell.' Though different styles and methods of knowledge engineering exist, the basic approach is the same: a knowledge engineer interviews and observes a human expert or a group of experts and learns what the experts know, and how they reason with their knowledge. The engineer then translates the knowledge into a computer-usable language, and designs an inference engine, a reasoning structure, that uses the knowledge appropriately. He also determines how to integrate the use of uncertain knowledge in the reasoning process, and what kinds of explanation would be useful to the end user.

Next, the inference engine and facilities for representing knowledge and for explaining are programmed, and the domain knowledge is entered into the program piece by piece. It may be that the inference engine is not just right; the form of knowledge representation is awkward for the kind of knowledge needed for the task; and the expert might decide the pieces of knowledge are wrong. All these are discovered and modified as the expert system gradually gains competence.

The discovery and cumulation of techniques of machine reasoning and knowledge representation is generally the work of artificial intelligence research. The discovery and cumulation of knowledge of a task domain is the province of domain experts. Domain knowledge consists of both formal, textbook knowledge, and experiential knowledge – the *expertise* of the experts [Feigenbaum, McCorduck and Nii, 1998].

Self Check Exercise

4) What is Knowledge Engineering?

Notes: i) Write your answer in the space given below.
ii) Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

16.3 ARTIFICIAL INTELLIGENCE BASED DECISION SUPPORT SYSTEMS

The development of an electronic decision support system requires the combined efforts of specialists of a particular subject, and must be developed with the cooperation of the end-users who use them. Specialists tend to be trained in rather narrow domains and are best at solving problems within that domain.

Decision support systems (DSS) are human–computer decision-making systems to support managerial judgements, and intuitions to solve managerial problems by providing necessary information, generating, evaluating and suggesting decision alternatives. Most organisational problems need a combination of quantitative and qualitative data processing.

Decision support systems (DSS) are a subset of computer-based information systems (CBIS) [Eom, 2001]. Management support systems consist of DSS, expert systems and executive information systems. In the early 1970s, scholars in the CBIS area began to recognise the important roles the information systems play in supporting managers in their semi-structured or unstructured decision-making activities. It was argued that information systems should exist only to support decisions, and that the focus of the information systems development efforts should be shifted away from structured operational control to unstructured critical decisions in organisations. Decisions are irreversible and have far-reaching consequences for the rest of the organizational life.

Self Check Exercise

5) What is DSS?

Note: a) Write your answer in the space given below.
b) Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

16.3.1 Evolution of Decision Support Systems (DSS)

Sean Eom [2001] traces the evolution of DSS as follows. The landmarks through which DSS have passed can be traced back to the TPS (Transaction Processing Systems) of 1950s. DSS were deemed as part of MIS and EIS systems, as knowledge based systems and further evolved as neural network based decision-making systems. Since the first electronic general-purpose computer was put into full operation in the early 1940s, data-processing techniques have been continuously advancing. It was in the late 1950s that many organizations began to utilise transaction processing systems (TPS) or electronic data processing (EDP) systems to automate routine clerical tasks such as payroll, inventory and billing. Further, DSS were seen as part of the MIS or management information systems. The 1960s witnessed the emergence of management information systems (MIS) with the development of database management systems for collecting, organising, storing and retrieving data. MIS were developed to extract valuable management information by aggregating and summarising massive amounts of transaction data and allowing user-interactive managerial queries. The inclusion of simple modelling and statistical methods as a component of MIS permits computer systems to make routine (structured) decisions.

It was not until 1970 that scholars began to recognise the important role computer-based information systems (CBIS) play in supporting managers in their semi-structured or unstructured decision-making activities. Since the 1970s, study of DSS has become an essential part of CBIS. The 1980s witnessed another wave of information technologies, the artificial intelligence-based expert systems (ES), which are to replace and mimic human decision makers in making repetitive decisions in a narrow domain. During the mid 1980s, executive information systems (EIS) emerged as an important tool to serve the information needs of executives. EIS provides timely and critical information, which has been filtered and compressed for tracking and control purposes. The latest addition to CBIS is artificial neural networks (ANN). Neural network computing involves building intelligent systems to mimic human brain functions. ANNs attempt to achieve knowledge processing based on the parallel processing method of human brains, pattern recognition based on experience, and fast retrieval of massive amounts of data. Fuzzy logic, genetic algorithm, and intelligent agents are some of the other intelligent techniques that can be used along with neural networks to improve the effectiveness of personal, group, and organisational decision-making.

16.3.2 Knowledge Based Decision Support Systems (Intelligent DSS)

Expert systems are otherwise referred to as knowledge based systems. Essentially these systems behave as experts based upon the knowledge that is built into it. Also behavioral specifications can be added in the inferencing stage to make the systems behave as an expert doctor or mineralogist as exhibited in some of the earliest expert systems.

According to Sean Eom [2001], an increasing number of systems are incorporating domain knowledge, modelling, and analysis systems to provide users the capability of intelligent assistance. Knowledge base modules are being used to formulate problems and decision models, and analyse and interpret the results. Some systems are adding knowledge-based modules to replace human judgments. Managerial judgments have been used to ascertain (assess) future uncertainty and to select assumptions on which decision models can be based.

Some decisions are both knowledge and data intensive. Consequently, a large amount of data usually requires considerable efforts for their interpretation and use. The knowledge-based DSS include a knowledge management component which stores and manages a new class of emerging AI tools such as machine learning and case-based reasoning and learning. These tools can obtain knowledge from prior data, decisions and examples (cases), and contribute to the creation of DSS to support repetitive, complex real-time decision-making. Machine learning refers to computational methods/tools of a computer system to learn from experience (past solutions), data and observations, and consequently alter its behaviour, triggered by a modification to the stored knowledge. Artificial neural networks and genetic algorithms are the most notable approaches to machine learning.

Consider a DSS in the field of agriculture. In agriculture, expert systems are capable of integrating the perspectives of individual disciplines (e.g. plant pathology, entomology, horticulture, agricultural meteorology) into a framework that best addresses the type of adhoc decision-making required of modern farmers. Expert systems can be one of the most useful tools for accomplishing the task of providing growers with the day-to-day, integrated decision support needed to grow their crops.

One of the attractive features of decision support systems is the program's ability to review a consultation and provide the user with an explanation for how its conclusion was derived. The explanation function is essentially a record of the

reasoning process used by the expert to resolve problems. It provides for a better understanding of how the conclusion was reached and instills in the user a greater confidence in the conclusion and the expert system. The accumulation of facts to be presented when an explanation is asked for is usually part of the development shell or software.

16.4 PATTERN RECOGNITION

Pattern recognition is widely used in different domains of human activities that vary from human genome studies, satellite image analysis for weather forecasting, fingerprint analysis for security reasons, speech recognition for remote applications, to optical character recognition for digitisation of documents. In the field of information processing and retrieval also some works have been done. Pattern Recognition is a process of extracting information and structure from a signal or image, by reference to known signals or images. It has an ability to recognise a given sub pattern within a much larger pattern. Alternatively, it can be said that a machine capable of pattern recognition can be trained to extract certain features from a set of input patterns. Pattern recognition belongs to the family of artificial intelligence and expert system. There are two approaches of pattern recognition. These are:

i) Statistical Pattern Recognition (based on statistical information extracted from the patterns)

ii) Syntactic Pattern Recognition (based on *a priori* knowledge)

A number of algorithms for pattern recognition has been developed that varies from application to application. Pattern recognition may use any of the two methods:

a) Unsupervised method (there is no known grouping of the examples)

b) Supervised method (there is known grouping of the examples in advance).

Figure 16.1 depicts the process and procedure of pattern recognition based on syntactic approach [Huang, 2002]. The process involves several separate transformation stages before a final decision is reached. Some stages consist of normalization (preprocessing), segmentation, feature selection, feature extraction, removing redundant or irrelevant information and appropriate classification. Table 16.1 shows some applications of pattern recognition in ISAR domains.

Table 16.1 : Some Applications of Pattern Recognition

Domain	Document Classification	Speech Recognition	Data Mining
Application	Internet search	Automatic operator	Finding meaningful pattern
Input	Text document	Speech waveform	Points in relational database
Classes	Semantic categories	Spoken words	Good cluster

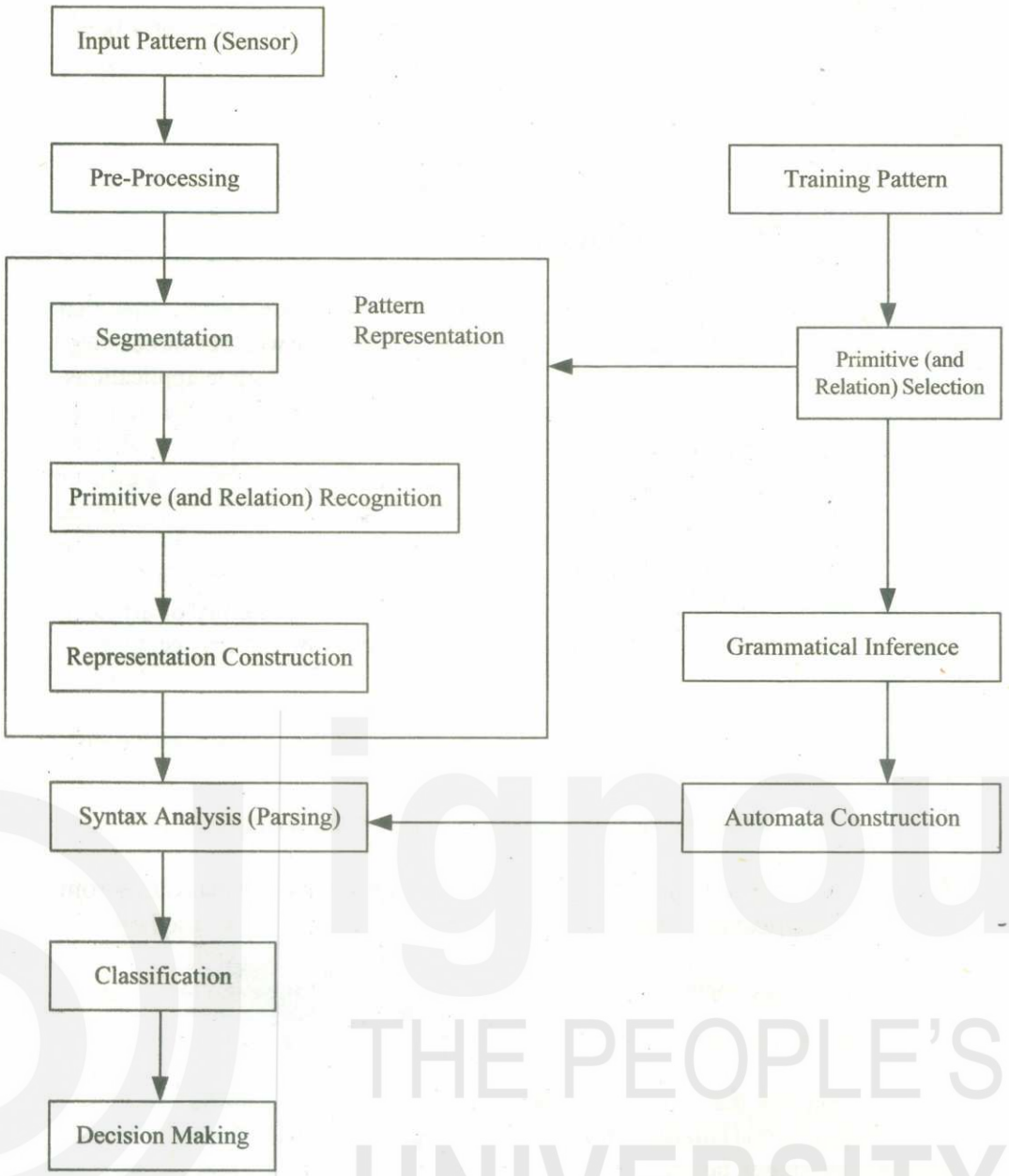


Figure 16.1 : Process of pattern recognition (syntactic)
(Source: www.worldscibooks.com/compsci/etextbook/4682/4682_chap1.pdf)

There are several attempts in recent times to use pattern recognition technologies in library and information services. Some possible applications of pattern recognition in library and information services are:

- *Library Security*: Movement of users can be monitored through recognizing the individual users.
- *Library Circulation*: Issuing of documents to the library patrons can be secured through recognizing the individual users.
- *Automatic Classification*: Attempts were undertaken for classifying the documents after analyzing patterns within the texts [Hamill and Zamora, 1978; McGill and Kowalski, 1978].
- *Document Search*: Intelligent modification of search terms (e.g. automatically correction of spelling mistakes, etc.) can be initiated through systematic categorisation.
- *Data Mining*: Users' information seeking behaviour and information usage pattern can be evaluated and future trends can be predicted analyzing library usage databases and statistics.

- *Optical Character Recognition* (OCR): Digitisation of library materials (rare books, theses, dissertations, manuscripts, journal articles, conference proceedings, etc.) is carried out in many libraries. OCR helps in converting an image of text into streams of characters that build up words and meaningful sentences. OCR software is available for English and some other European languages. Various OCR software for Indian languages are in development stage that will provide solutions for digitising our indigenous knowledge resources.

- *Speech Recognition*: Interactive voice-based library and information services are being provided in some libraries and information centres using speech recognition technologies.

Pattern recognition is widely used in the advanced and complex information retrieval systems, mostly in the areas of biochemical, pharmaceutical and medical information. For example, Chemical Abstracts can be searched by molecular structure of biochemical compounds. Similarly, ISAR systems of human genomes can be searched by DNA protein sequence. There are some ISAR systems of fine arts that use pattern recognition techniques for retrieval of information on paintings, drawings, illustrations, designs, sculptures, and other objects [Aleksander, 1982]. Some other ISAR systems, meant for criminal investigation, use pattern recognition techniques for identification of persons based on their fingerprints, facial patterns, etc.

16.5 SUMMARY

Information retrieval is an intellectual task as it involves query analysis and efficient matching and retrieval of concepts. Expert Systems is one area of Artificial Intelligence that has been most result oriented and several systems have been successfully designed to emulate the experts' reasoning process in several diverse fields. Decision Support Systems based upon Artificial Intelligence techniques have been developed for specialised areas such medical diagnosis and agriculture.

16.6 ANSWERS TO SELF CHECK EXERCISES

1) The study of AI itself serves as means of understanding human intelligence. The reasons stated are as follows:

   - The use of computers demands a clear statement of the problem and clear strategy for solutions and this requires a clear understanding about the human thinking process.
   - Computer models force precision. Implementing a theory uncovers conceptual mistakes and oversights that ordinarily escape the notice of even the most meticulous researcher.
   - Computer implementations qualify task requirements. Once a program performs a task, upper bound statements can be made about how much information processing the task requires.
   - It is usually simple to deprive a computer program of some piece of knowledge in order to test how important that information really is.

2) In general, expert systems are composed of basic components such as:

   i) a user interface: To facilitate user interaction

   ii) a knowledge base: The facts or knowledge based upon which the ES makes decisions

   iii) an inference mechanism: the reasoning engines built according to heuristics reasoning or facts.

3) The *knowledge base* of expert systems contains both factual and heuristic knowledge:

   Factual knowledge is that knowledge of the task domain that is widely shared, typically found in textbooks or journals, and commonly agreed upon by those knowledgeable in the particular field.

   Heuristic knowledge is the less rigorous, more experimental, more judgmental knowledge of performance. In contrast to factual knowledge, heuristic knowledge is largely based upon past experiences and knowledge gained. For instance the knowledge that, if the same pattern of events take place then the same conclusion could be expected. It is the knowledge built from 'thumb rules'.

4) For computer systems to behave as an 'expert system', expertise has to be built into the system. Knowledge Engineering refers to techniques of building knowledge into expert systems. It involves understanding knowledge in specific domains and representing it in computer in a syntax that is computer process-able.

5) Decision support systems (DSS) are human–computer decision-making systems to support managerial judgments, and intuitions to solve managerial problems by providing necessary information, generating, evaluating and suggesting decision alternatives. Most organisational problems need a combination of quantitative and qualitative data processing. Decision support systems (DSS) are a subset of computer-based information systems (CBIS). DSS maybe part Management support systems or also expert systems and executive information systems.

16.7 KEYWORDS

AI : Artificial Intelligence – deals with techniques of making intelligent systems

CBIS : Computer Based Information Systems

Inference Engine : A component of expert systems that makes decisions based upon reasoning built into the systems through factual or ruled based methods

KBS : Knowledge Based Systems, also known as Expert System and Knowledge Based Computer Systems (KBCS)

Knowledge Base : Facts or knowledge of a specific domain that is built into an expert system based on which the system can make decisions

Knowledge Engineering : Techniques of building knowledge into expert systems. That is understanding knowledge in specific domains and representing it in computer in a syntax that is computer processible

16.8 REFERENCES AND FURTHER READING

Aleksander, I. (1982). Modern pattern recognition and the classification of works of art. *Art Libraries Journal*, 7(2), 61-66.

Borrow, D.G. and Winoguard, J. (1977). An overview of KRL, a knowledge representation language. *Cognitive Science*, 1, 3–46.

Brajnik, Guida and Tasso. (1986). *An expert interface for effective man-ma-*

chine interaction. In: *Cooperative Interfaces to Information Systems*. L.Bolc, M.Jarke (eds.). pp. 259 – 308. Berlin: Springer-Verlag.

Carson-Berndsen, J and Gazdar, G. *Lexcical knowledge representation.* <http://www.ccl.kuleuven.ac.be/LKR/html> (browsed on 21 July 2004).

Decision support system. <http://cstl-hcb.semo.edu/eom/iebmdssrwweb.PDF>.

Engelmore, Robert S. and Feigenbaum, Edward. *Report of the JTEC panel on knowledge-based systems in Japan*. <http://www.wtec.org/loyola/kb/toc.htm> (accessed on 11th July 2004).

Eom, Sean B. (2001). *Decision support system*. In: *International encyclopedia of business and management*. Malcolm Warner (ed.). 2nd ed. London: International Thomson Business Publishing.

Feigenbaum, E.; Mccorduck, P. and Nii, H.P. (1998). *The rise of the expert company*. New York: Times Books.

Feigenbaum, F.A. *Art of artificial intelligence: themes and case studies of knowledge engineering*. In: *International Conference of Artificial Intelligence*. Cambridge: Massachusetts Institute of Technology.

Gottlieb, T.Z. and Boulton, P.I.P. (1975). *A pattern recognition document classification technique*. In: *Third open conference on information science in Canada*, Quebec, 8-10 May, 1975. Ottawa: Canadian Association for Information Science.

Hamill, Karen A. and Zamora, Antonio (1978). *An automatic document classification system using pattern recognition techniques*. In: *The information age in perspective: proceedings of the ASIS Annual Meeting.* New York: Knowledge Industry Publications.

Hayes-Roth, Fredrick, Waterman, Donald and Lenat Douglas, B. (1983). *An overview of expert systems*. In: *Building expert systems*. London: Addison-Wesley.

Huang, Kou-Yuan. (2002). *Syntactic pattern recognition for seismic oil exploration.* Singapore: World Scientific Publishing.

King, D. (1993). *Intelligent support systems: art, augmentation, and agents*. In: *Decision Support Systems: Putting Theory into Practice*, edited by R.H. Sprague, Jr. and H.J. Watson. 3rd ed. Englewood Cliffs, NJ: Prentice Hall.

Leibowitz, J. (1990). *Dynamics of decision support systems and expert systems*. Chicago: Dryden.

McGill, James R. and Kowalski, B.R. (1978). Classification of mass spectra via pattern recognition. *Journal of Chemical Information and Computer Sciences*, 18(1), 52-55.

Meredith, J.C. (1971). Machine-assisted approach to general reference materials. *Journal of the American Society for Information Science*, 22 (3), 176-186.

Minsky, M.A. (1975). *Frame for representing knowledge*. In: *Psychology of Computer Vision*, edited by P.H. Winston. New York: McGraw-Hill.

Newell, A. (1969). *Heuristics programming: III structured problems*. In: *Progress in operation research*. Volume 3. New York: John Willey.

Patterson, D.W. (1992). *Introduction to artificial intelligence and expert systems*. New Delhi: Pretice-Hall of India.

Politt, A.S. (1986). *A rule – based system as an intermediary for searching cancer therapy literature on MEDLINE*. In: *Intelligent Information System: Progress and Prospects*. pp. 82 – 126. Chichester: England Ellis Horwood.

Politt, A.S. (1987). CANSEARCH: an expert system approach to document retrieval. *Information Processing and Management*, 23 (2), 100–108.

Rich, Elaine. (1990). *Artificial intelligence*. 2nd ed. New York: McGraw-Hill.

Stuper, A.J. and Jurs, P.C. (1976). ADAPT: a computer system for automated data analysis using pattern recognition techniques. *Journal of Chemical Information and Computer Sciences*, 16(2), 99-105.

Waterman, D.A. (1986). *A guide to expert systems*. Massachusetts: Addison-Wesley.

Weil, C.B. (1968). Automatic retrieval of biographical reference books. *Journal of Library Automation*, 1(4), 239 – 249.

What are expert systems? <http://esdg.cas.psu.edu/default.html> (Browsed on 11th August 2004).

Winston, Patrick Henry. (1984). *Artificial intelligence*. 2nd ed. London: Addison–Wesley.

Zarri, Gian Piero. (1985). Interactive information retrieval: an artificial intelligence approach to deal with bibliographic data. *Advances in Intelligent Information Retrieval: Informatics*, 8, 101-119.

Zeleny, M. (1987). Management support systems: towards integrated knowledge management. *Human Systems Management*, 7(1), 559-70.