ignou
THE PEOPLE'S
UNIVERSITY

Indira Gandhi
National Open University
School of Social Sciences

BLI-223
Organising and Managing Information

Block

4

RECENT DEVELOPMENTS

Programme Design Committee

Prof. Uma Kanjilal (Chairperson)
Faculty of LIS, SOSS, IGNOU

Prof. B.K.Sen, Retired Scientist, NISCAIR
New Delhi

Prof. K.S. Raghavan, DRTC
Indian Statistical Institute, Bangalore

Prof. Krishan Kumar, Retired Professor
Dept. of LIS, University of Delhi, Delhi

Prof. M.M. Kashyap, Retired Professor
Dept. of LIS, University of Delhi, Delhi

Prof. R.Satyanarayana
Retired Professor, Faculty of LIS, SOSS, IGNOU

Dr. R.Sevukan (Former Faculty Member)
Faculty of LIS, SOSS, IGNOU

Prof. S.B. Ghosh, Retired Professor
Faculty of LIS, SOSS, IGNOU

Prof. T. Viswanathan, Retired Director
NISCAIR, New Delhi

Dr. Zuchamo Yanthan
Faculty of LIS, SOSS, IGNOU

Conveners:

Dr. Jaideep Sharma
Faculty of LIS, SOSS, IGNOU

Prof. Neena Talwar Kanungo
Faculty of LIS, SOSS, IGNOU

Programme Coordinators

Prof. Jaideep Sharma and
Prof. Neena Talwar Kanungo

Course Coordinator

Prof. Jaideep Sharma

Programme Editor

Prof. Jaideep Sharma

Course Preparation Team

Unit No(s).	Contributor(s)
12-13	Dr. Asok Mukhopadhya
14	Dr. Aditya Tripathi

Internal Faculty:
Prof. Jaideep Sharma
Prof. Neena Talwar Kanungo

Course Editor
Prof. Jaideep Sharma

Material Production

Mr. Manjit Singh
Section Officer (Pub.)
SOSS, IGNOU

Secretarial Assistance

Ms. Sunita Soni
SOSS
IGNOU

Cover Design

Ms. Ruchi Sethi
Web Designer
E Gyankosh
IGNOU

November, 2013

ISBN :

"The University does not warrant or assume any legal liability or responsibility for the academic content of this course provided by the authors as far as the copyright issues are concerned"

Further information on Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi, New Delhi-110 068 or visit University's Website http://www.ignou.ac.in.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by Director, SOSS.

Lasertypesetted at Graphic Printers, 204, Pankaj Tower, Mayur Vihar, Phase-I, Delhi-110091.

Printed at :

BLOCK 4 RECENT DEVELOPMENTS

Introduction

This Block is devoted to the developments taking place in organising and managing information. Knowledge organisation, as a bibliographic control oriented concept, has a long history. Originating in ancient time, it has developed in stages reflecting changing societal information demands. It began as a straight forward list of items, often without any order. Later it took the shape of inventory lists for recording the stock against serial numbers followed by adding the location marks for retrieval. Cataloguing as collocating device was introduced in the 19th century. From the second half of the 20th century information technology has taken the lead, backed by computerised database, the Internet technology, Web resources, and new Metadata tools, in achieving the mission and objectives of Knowledge Organisation. Today, in the 21st century a social dimension has been added to take care of the personal informational requirements of individuals, besides the collective demands of the human societies.

Unit 12, titled **Conceptual Changes: Impact of Technology** is spread over six different topics touching upon different aspects of Knowledge Organisation Systems (KOS). These include: Origin and types of KOS, Analysis and planning of KOS, Methods of linking dispersed digital resources, Methods of accessing heterogeneous networked resources, Contributions of W3C in developing standards like Description Framework (RDF), Web Ontology Language (OWL), and particularly, Simple Knowledge Organisation Systems (SKOS). Lastly, the future of semantic activities is discussed in the context of some identified problems yet to be resolved.

Unit 13 **Online Cataloguing: Design and Service** begins by reviewing the development of library catalogue from its early stage up to its latest manifestation in networking environment. It is followed by a discussion on cataloguing standards. The internal structure of a MARC record is examined taking the requirements of bibliographic fields in view. Next, we observed the functionality of few metadata tools in building online catalogues, including AACR2/RDA. Then, OPAC as user interface is discussed in the context of advancing Web 2.0 technologies. Towards the end the requirements of 'copy cataloguing' and 'original cataloguing' and the features of various utilities, offered by network services for generation, conversion, validation of MARC records, and other supports have been explained in detail. This gives an exposure to the different tasks associated with cataloguing of networked resources, apart from awareness of the utility services as such.

The last Unit, no. 14 in this Block is **Overview of Web Indexing, Metadata, Interoperability and Ontologies**. The concept of web indexing and its importance has been discussed in the Unit followed by a description of its types. The concept of metadata has been explained, delineating its different types. Interoperability is very important keeping in view the disparities in the systems used and their frequent interaction and sharing of information with each other. The Unit is devoted to a discussion on its need, methods and protocols for achieving interoperability.

UNIT 12 CONCEPTUAL CHANGES: IMPACT OF TECHNOLOGY

Structure

12.0 OBJECTIVES

After reading this Unit, you will be able to:

- express the meaning and purposes of Knowledge Organisation Systems (KOS);
- explain the analytical principles for organisation of the intellectual content of information records;
- describe the methods of creating and providing access to records; and
- organise knowledge and information resources meaningfully and purposively from the perspective of a networked environment.

12.1 KNOWLEDGE ORGANISATION: AN OVERVIEW

Knowledge Organisation (KO) is a field of investigation closely related to Information Retrieval (IR). An introductory discussion on KO demands a clear understanding about knowledge, information and data, and their hierarchic relationship.

12.1.1 Knowledge Hierarchy

Information is one of the most frequently used words in our society, and widely differs in its meaning – from news to intelligence, data to knowledge. However, in the context of library and information science, information is understood as something more than data. Data is unrelated whereas information is related.

Information reflects an organisation of data. Characteristically, it provides verifiable statements of facts that can be either true or false. Knowledge, on the other hand, belongs to a higher plane of organisation, assumes a verified status of truth or falsity, and comprehends generalised pattern of information.

The following diagram represents the transitions from data, to information, to knowledge, and ultimately to wisdom. Transition from a lower stage to higher up is carried on through 'understanding', which has no separate level of its own. The structural and functional relationships between data, information, knowledge, and wisdom are represented in the following Data-Information-Knowledge-Wisdom (DIKW) model:

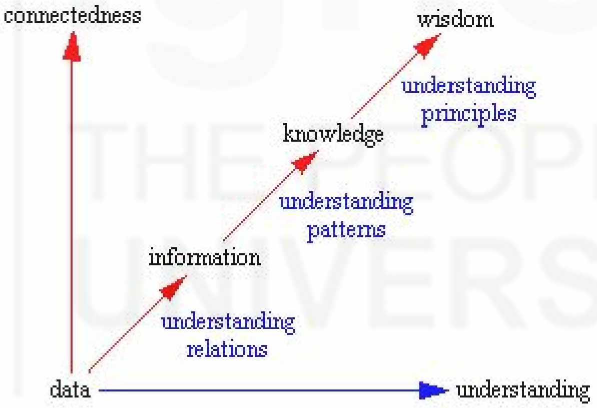


Fig. 12.1: DIKW model

Source: Gene Bellinger http://www.systems-thinking.org/dikw/dikw.htm

12.1.2 Knowledge Organisation : Concept

Knowledge Organisation, often referred to as KO, is a field of investigation within Library and Information Science (LIS). As already mentioned, KO is directly related to Information Retrieval (IR) – the science of searching for documents, for information within documents and for metadata about documents. KO investigates into the nature and order of knowledge, and primarily concerned with *grouping of like things*, documents, for information within documents and for metadata about documents. KO investigates into the nature and order of knowledge, and primarily concerned with *grouping of related things*.

In context of LIS, the word, *things,* may mean:

- Physical documents, or

- Parts of the documents, or
- Representations (Surrogates) of the documents, or
- Representations (Surrogates) of the parts of the documents, as well as
- Concepts (Metadata) used to characterise documents.

Any grouping of objects is based on conceptualisations of the things to be classified.

Example

Suppose we classify *Peacock* as 'National birds', *and not* as 'Peafowl'. In that case, documents about peacocks are grouped with all other documents on National birds, of variant species. Moreover, following the order of knowledge organisation, we may expect to find groupings on national animals, national flowers, and other national *things* in the neighbourhoods.

The example shows:

- The distinction between classifying *things* and classifying documents or concepts is of little significance theoretically. By implication, KO is about the grouping of related *concepts*.
- Our worldview is always dependent on our pre-understandings and conceptualisations of the objects in the world (e.g. National birds, National flowers). We interact with our world with pre-conceived ideas and objects.
- Conceptualisation of things is a culture-specific process, dependant on the cultural outlook and domains of the people. This implies that people from different walks of life with different cultural background tend to classify things differently.
- From socio-cognitive perspective, concepts are *shared classifications.* When people imbibe a culture, learn a language, or embrace a profession, they inherit some common outlook for classifying some part of the world, e.g. public transports.
- Concepts facilitate user interaction with natural and cultural environment. Concepts are generally functional, and they serve pragmatic ends.
- The important question from KO perspective is – which concepts should be preferred in attaining the goal of the information system.
- The success or failure in attaining the goal rests on the selection of controlled vocabularies, and classifications for a certain concept (or meaning) and repression of the alternative concepts/ meanings/ views.

Knowledge Organisation Helps Information Retrieval

As discussed already, Information Retrieval (IR) and KO are closely related fields within LIS. IR relies heavily on some form of KO. It is interesting to note that we create some KO when information is stored (e.g. Class Number, Book Number, Subject Headings), and some KO is created on the fly. For instance, computer information retrieval generates frequency of usage of the search terms that helps to search for information. IR is an intellectual device for providing access to anything stored anywhere, yet the order of which might not be evident. KO, by navigating natural orders, and creating and imposing useful orders helps IR to reach its goal.

Self Check Exercises

Note: 1) Write your answers in the space given below.

2) Check your answers with the answers given at the end of this Unit.

1) What is the continuum of data, information and knowledge? How those are mutually related?

....

....

....

....

2) What is a KOS? What are the common characteristics of KOSs?

....

....

....

....

12.2 KNOWLEDGE ORGANISATION SYSTEMS: ORIGIN AND TYPES

12.2.1 Knowledge Organisation in the Pre-Digital Age

Knowledge organisation, as a bibliographic control oriented concept, has a long history. Knowledge organisation originated in ancient time. Vedas (c 500-400 BCE) is perhaps the oldest specimen of categorisation of human knowledge in four categories. In the Han Dynasty, a library classification system was formulated. In the 19th century, Thomas Jefferson devised his classification system on Baconian methods. Anthony Panizzi's cataloguing principles and Charles Ammi Cutter's Rules for Dictionary catalogues culminated in Paris Principles, and AACR2.

Hodge (2000) pointed out that there has been always a need in a traditional library situation to store a single item at a single location on a shelf. All the bibliographic classification schemes, like LC, DDC, CC, and UDC were developed to serve this purpose. To provide multiple access points libraries use subject heading schemes such as LCSH, Sears, or other specialised schemes at micro level for specific subject areas. Libraries developed Authority Files to control variant form of personal, organisational, and geographic names for searching and browsing local collections.

All these semantic schemes, initially constructed to control print-dominated resources, made enormous contribution in developing Knowledge Organisation Systems to make the networked resources accessible.

Five Main Stages in History of Knowledge Organisation

By examining the trends in scholarly persuasions under the socio-economic and technological influences, we have traced the stages of KO's development with their characteristic features.

- Antiquity: Lists

- Middle Ages: Inventories
- Seventeenth Century: Finding lists
- Nineteenth Century: Collocating devices
- Twentieth Century: Automation and Codification
- Twenty-first Century: Social and Collaborative Tagging

Each represents a turning point that reflects changing societal information demands and the development of new technologies. At the beginning, there were only straight forward list of items, often without any order. During the Middle Ages, making of inventory lists were introduced for recording the stock against serial numbers. Then in the next phase, the location marks were added to the inventory for retrieval. Cataloguing as collocating device was introduced in the 19th century. From the second half of the 20th century information technology has taken the lead, backed by computerised database, the Internet technology, Web resources, and new Metadata tools, in achieving the mission and objectives of Knowledge Organisation. Today, in the 21st century a social dimension has been added to take care of the personal informational requirements of individuals, besides the collective demands of the human societies.

Knowledge Organisation: More than Bibliographic Control

Today the term 'Bibliographic Control' is gradually fading out for more than one reason. First, it stems from a print-dominated book-oriented world. Second, the conventional bibliographic tools are losing their relevance to networked resources. Though, KOs and Bibliographic tools are essentially same, they are different in certain respects:

- Knowledge Organisation is much broader than simply bibliographic control.
- Knowledge Organisation is concerned with understanding how knowledge is generated and used.
- Such knowledge helps us employ relatively more sophisticated approaches to information retrieval.

Knowledge Organisation Systems (KOS)

KOS refers to the semantic tools that present the organised interpretation of knowledge structures. In this broad sense, libraries, encyclopaedias, academic disciplines and such other knowledge organisation systems may serve as examples of KOS. However, it is all important for the development of knowledge organisation, as an intelligent system, to know how far the traditional semantic tools and schemes are relevant and effective in networked resource environment.

The term 'Knowledge Organisation Systems' was coined by *the Networked Knowledge Organisation Systems Working Group* at *the ACM Digital Libraries 098 Conference* in Pittsburgh, Pennsylvania. KOS does not include anything more than what KO does, other than its emphasis on *system*. In a general way, KOS refers to the semantic tools that present the organised interpretation of knowledge structures. In this broad sense, libraries, encyclopedias, academic disciplines and such other knowledge organisation systems may serve as examples of KOS.

Gail Hodge (2000), one of the renowned exponents of KOS, writes:

"The term, *knowledge organisation systems,* is intended to encompass all types of schemes for organising information and promoting knowledge management. Knowledge organisation systems include classification and categorisation schemes that organise materials at a general level, subject headings that provide more detailed access, and authority files that control variant versions of key information such as geographic names and personal names. Knowledge organisation systems also include highly structured vocabularies, such as thesauri, and less traditional schemes, such as semantic networks and ontology. Because knowledge organisation systems are mechanisms for organising information, they are at the heart of every library, museum, and archive."

Common Characteristics of Knowledge Organisation Systems

KOS imposes a particular view of the world.

- The same entity can be characterised in different ways depending on the KOS that is used.
- There must be a sufficient commonality between the concept in KOS and the real world objects it refers.
- A person seeking relevant material by using a KOS must be able to connect his or her concept with its representation in the system.

KOS imposes a particular view of the world on a particular collection through:

- Providing a controlled list.
- Controlling synonyms or equivalents.
- Linking DL (digital library) resources to related resources.
- Making semantic relationships explicit.

Knowledge Organisation Systems for Digital Libraries

- KOS is intended to encompass all types of schemes for organising information and promoting knowledge management.
- Includes traditional classification schemes, subject headings, thesauri, etc.
- Includes less traditional schemes such as semantic networks and ontology
- All digital libraries use one or more KOS.

Hodge points out that there can not be a single knowledge classification scheme on which everyone agrees. A single KOS would have been advantageous, if ever be developed. Cultural diversity makes such an ideal KOS unattainable. This is because of the difference in cultural values and social attitudes, something considered meaningful to one community, not necessarily meaningful to another. Therefore, we live in a world of multiple, variant ways to organise knowledge. (Lesk 1997)

Despite their diversity, we can identify some common characteristics of KOS critically important for their use in organising digital libraries.

Knowledge Organisation for Web Resources

Simple Knowledge Organisation System, or SKOS, provides a model representing the basic structure and the content of *concept schemes*, which may be thesauri, classification schemes, and subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary. SKOS is an application of the Resource Description Framework (RDF), and it allows concepts to be composed and published on the World Wide Web, linked with data on the Web and integrated into other concept schemes.

12.2.2 Knowledge Organisation Systems: Types

There are different systems possible for orgainising digital libraries. Given below their descriptions are based on characteristics such as structure and complexity, relationships among terms, and historical function. KOS are grouped into three general categories:

Term Lists

Authority Files

Authority Files are lists of terms that are used to control the variant names for an entity or the domain value for a particular field. Examples include names for countries, individuals, and organisations. The presentation may be alphabetical or organised by a shallow classification scheme. Specific examples of authority files include the Library of Congress Name Authority File, Sears List of Subject Headings, and the Getty Geographic Authority File.

Glossaries

A glossary is a list of terms, usually with definitions. The terms may be from a specific subject field, or those used in a particular work. Examples include the EPA Terms of the Environment, ALA glossary of library and information science.

Gazetteers

A gazetteer is a dictionary of place names. Traditional gazetteers have been published as books or they appear as indexes to atlases. Each entry may also be identified by feature type, such as river, city, or school. An example is the Geographic Names Information Service, The Columbia Gazetteer of the World.

Dictionaries

Dictionaries are alphabetical lists of terms and their definitions that provide variant senses for each term, where applicable. A dictionary also provides sometimes synonyms and, through definitions, related terms. There was, however, no explicit hierarchical structure or attempt to group terms by concept.

Classification and Categorization

Subject Headings

Subject heading lists can be extensive, covering a broad range of subjects. However, the structure of these lists is generally shallow, with limited hierarchy. The subject headings are pre-coordinated, with rules for constructing headings narrowing down the scope of a concept. Examples include the Medical Subject Headings (MeSH) and the Library of Congress Subject Headings (LCSH).

Classification Schemes, Taxonomies

In general, these types of KOSs provide ways to separate entities into *buckets* or relatively broad topics. Some examples provide a hierarchical arrangement of numeric or alphabetic notations to represent broad topics. These types of KOSs lack the explicit relationships presented in a thesaurus. Examples of well-liked classification schemes include LCCS, DDC, and UDC. Subject categories are often used to group thesaurus terms in broad topic sets, outside the hierarchical scheme of the thesaurus.

Taxonomies are increasingly being used in object oriented design and knowledge management systems to indicate any grouping of objects based on a particular characteristic.

Relationship Group

Thesauri

These KOSs are based on concepts, and they show relationships between terms. Relationships commonly expressed in a thesaurus include hierarchy, equivalence, and associative (or related). These relationships are generally represented by the notation of broader term, narrower term, synonym, and associative or related terms. Associative Relationships may be more granular in some schemes. For example, the Unified Medical Language System (UMLS) provides over 40 associative relationships. Entry terms point to the preferred terms that are to be used for each concept. Most were developed for a specific discipline, for example, FAO Aquatic Sciences and Fisheries Thesaurus, NASA Thesaurus for aeronautics and aerospace-related topics.

Very recently, a new kind of thesauri has come up on the Web named *Visual Thesaurus*, which allows you to discover the connections between words in a visually captivating display.

Semantic Networks

Semantic network is one of the most significant developments in the area of natural language processing technology. Here the concepts and terms not positioned as hierarchies but as a network or a Web. Concepts are thought of as nodes with various relationships branching out from them. These include specific relationships of whole-part, cause-effect, parent-child, etc, instead of standard BT, NT and RT. Princeton's WorldNet is a grand example of semantic network.

Ontologies

In general, ontology is the study or concern about what kinds of things exist – what entities there are in the universe. In information technology, ontology is the working model of entities and interactions in some particular domain of knowledge or practices, such as electronic commerce or "the activity of planning." We may define Ontology as specification of conceptualisations, used to help programs and humans share knowledge. Ontologies are being developed as specific concept models by the Knowledge Management community.

Self Check Exercises

Note: 1) Write your answers in the space given below.

2) Check your answers with the answers given at the end of this Unit.

3) Knowledge Organisation Systems have evolved in stages. Describe those stages highlighting their characteristics.

....

....

....

....

4) Name the types of Knowledge Organisation Systems. State briefly what you understand about Semantic Networks and Ontologies.

....

....

....

....

12.3 PLANNING KNOWLEDGE ORGANISATION SYSTEMS

12.3.1 Analysing User Needs

For any type of library, print-dominated or digital, process of planning starts from assessment of users' needs in terms of content and functionality. Everything else, in terms of resources, and service programs, grows out of the users' needs. The digital library is a *control zone*. It means that a digital library " is always a highly selective subset of available information objects, segregated and favoured, to which access is enhanced and to which the attention of client-users is drawn in opposition to objects excluded." (Atkinson 1996)

An important characteristic feature of a digital library is that it has in-house collections of its own and unbound networked resources within its reach. Therefore, apart from the needs for organising the in-house collections, planning must be done for establishing possible links between contents within and outside the digital library walls. Since KOSs are to act as intermediate authority files, links must always show their relationship to user needs.

12.3.2 Locating Knowledge Organisation Systems

After mapping the user needs decisively, the next task is to locate suitable KOSs, if any that exist. It is preferable to use an existing KOS. Because, building KOSs is costly and time-consuming proposition. The value of a KOS comes from its user community when they appreciate it. By creating a control zone, that is, by selecting some objects and excluding others, information professionals are using their expertise to bring users to documents that hold a particular value. The selection policy must acknowledge the users' behaviour. In users' assessments, the sources built by learned societies, professional associations, or standards groups, stand out more trustworthy than those built in-house.

We must recognise the fact that the networked environment has resulted in both an explosion of primary materials, including documents, electronic journals, and databases, and in an equivalent explosion of KOSs on the Web.

12.3.3 Planning the Infrastructure

Closed Zone

In the context of the digital library, the physical location of the KOS is a critical factor for deciding upon the architecture of KOS. It is on the physical location the position of KOS depends – whether it will be held externally or internally. Both have their own advantages and disadvantages. If KOS is held internally, that is in controlled zone; classification, thesaurus, and ontologies need to be hand-picked cautiously by applying selection criteria, professional expertise and users' feedback. We should safeguard against all possible risks due to the following:

- Selection might limit the access.

This may happen due to various reasons; e.g., when a less inefficient classification scheme is chosen for shelving arrangement, when class numbers are pre-coordinated leaving little scope for libraries to accommodate new concepts; when a chosen subject heading scheme fails to provide unaffected distribution of concepts, e.g. disproportionate distribution of world religions, or other cultural bias; when primary search terms and their spellings are not guided by international standards; and the like shortcomings of KOSs.

- Selection might be biased.

  This may happen when there is no well-documented articulate selection policy to follow; or when standard selection criteria are ignored to give preference to a biased judgment.

- Knowledge organisation supports pre-existing concepts, not for new concepts.

  The traditional semantic tools are inflexible and restricted to the limit of possibilities they have had at starting. It must be ensured that KOSs have in-built mechanisms to accommodate new concepts, terminology, and even search features to perform unaffectedly.

- Not "user-oriented" – individual user's needs are different.

  This may be looked upon as most unfortunate failure from the view point of prevailing Library 2.0 principles where importance of individual requirements of a user have been duly stressed. Libraries must take cognition of the personal needs as well as the collective needs of their users while planning for KOSs.

  On the positive side, the KOS is under more local control. Therefore, it may be possible to improve the response time by not accessing the KOS over the Internet. If the KOS is to be used behind the scenes, concerns of speed and integration become more important. If additional modifications (including digitisation) need to be made to the KOS to integrate it with the digital library, it will also be necessary to load the KOS locally.

Open Zone

If the system is available in open zone on the Web, KOS as an external system, its architecture requires a script to locate the resource. One must then launch a query against the resource to obtain the piece of information that will serve as the *key* between the two files. This key could be

- A universal resource locator (URL) or
- Input to another search query.

A query may be necessary if the KOS is stored in a database. The script may transfer log-on information (including user ID and password) from the digital library system to the external KOS, in order to provide access to the Web-enabled database. In the case of a more direct link, the access may be by URL. Another alternative is the use of other Uniform Resource Identification (URI) schemes and the Uniform Resource Name (URN), which can be sent from the newer Web browsers.

The benefits of linking to a remote resource are that:

- The resource will always be up-to-date.
- The maintenance of the KOS is in the hands of the owner, not the digital librarian.

- It may also be more apparent to users that the KOS is not owned by the digital library.

12.3.4 Change, Upgradation and Version Updation

Like any social or intellectual organisation, an unexpected change in the organisation and content of the system poses problems. Therefore, there should necessarily be a planning process to continue behind the scene to predict and implement changes systematically. Few areas warrant constant review. The software or telecommunications route between the digital library server and the KOS may prove unreliable. The KOS may be obtained from the owner and found tricky when loaded locally. In many cases, this requires licensing that may not be required when the KOS is accessed remotely. Loading a KOS locally also involves issues such as maintenance, local system administration, and disk storage. If the KOS uses special software, such as a database management system, loading the KOS locally will require a copy of that software, which may involves added cost toward purchase or licensing. Besides KOS installation, there are more systems related problems, like firewalls and interface design.

For a digital library, an outdated KOS *can be more of a hindrance than a benefit*. When planning for installation, issues related to maintenance of the KOSs should be settled and blue printed. Version control of the KOS is extremely important. Reloading a new version from the system provider is one way to accommodate changes; however, this may not be acceptable if the locally held version differs substantially from the one currently being in operation. If there has been a significant transformation through customisation procedures, it may prove difficult, if not impossible, to reload the original and recreate the changes that have been made. There is also a marked drift toward developing platform-dependent software. Sometimes, platform changes more frequently than does the software.

There is yet another way of updating KOSs, which is known as *transaction-based approach*, whereby only changed components are transferred by the KOS provider to the library. This, however, requires that the system provider have the necessary infrastructural facilities to effect these transactions on time. In fact, this transaction-based approach is becoming the favourite among version updating options. We now find increasing number of KOS publishers regularly report changes that have taken place since the previous version issued. However, the changes are often not indicated with enough detail to support automatic change transactions.

12.3.5 Presenting the Knowledge Organisation System to the User

Textual Representation

KOS architecture must accommodate the character sets of the incoming sources. This is particularly important if a data string, in ASCII or mark-up language, has been used to represent special characters and diacritical marks. Systems that have been developed in Unicode, which generously accommodate all the existing scripts of languages worldwide have decisively removed the age-old constraints of inter-lingual communications. With the support of Unicode standards, all electronic and web-based KOS platforms are now empowered to communicate directly in original language scripts, that is, without resorting to diacritical marks – a clumsy means of phonetic representations based on extended ASCII. Although the Unicode is now universally accepted standard in the software industry, we still find instances of exceptions. Therefore, the Unicode compliancy needs to be ensured even today when going for a new KOS.

Visibility and Presentation

From the user point of view, the ways a KOS presents itself is no less important than its content is. Because, an ill-designed view front fails to convey the content meaningfully, Moreover, a poor presentation may discourage a person to use the KOS. Therefore, while deciding which KOS should be used and what functions it should serve, the digital library will need to determine how to present the KOS to its users.

The KOS can be exposed to the user in different ways. In the website of the digital library, KOS-related themes or categories may be clustered logically for users' view and interactions. The KOS may be used at a higher level to identify specific portals launched by different user communities. If the content of the digital library includes metadata records, the KOS may be displayed as index terms on the records or as an independent navigation tool. KOS may also remain transparent. For example, a database search procedure may employ a thesaurus behind-the-scene to dig up synonyms for using as multiple search keys. This way, KOSs make presentations of real time statistical information keeping the calculation system behind the scene.

Self Check Exercises

Note: 1) Write your answers in the space given below.

2) Check your answers with the answers given at the end of this Unit.

5) What are the possible risks of Knowledge Organisation Systems in Controlled zone? Suggest safeguards against the risks?

....

....

....

....

6) Determine how to present the Knowledge Organisation Systems to its users.

....

....

....

....

12.4 LINKING INTERRELATED DIGITAL RESOURCES

12.4.1 Relations

Efficacy of KOSs largely depends on their ability of interlinking related digital library resources. KOSs are used for linking digital resources to other digital resources or, indirectly, to physical objects. We may define linking in many ways. We can say in plain words that:

A link is a connection from one page to another destination such as another page, or a different location on the same page.

By means of linking device, identification and categorisation of relations are achieved, which is "a necessary requirement for the formal description that makes navigation possible in the bibliographic universe" (Husby 2001).

The relations may be one of these kinds:

a priori – given by the nature of things. A link is an expression of a relation.

There are many ways of expressing relation. Not all are hypertext links, and some follow quite different methods, such as:

- Citing together

  An article cites more than one item; the cited items are related with each other in reference to the citing source.

- Explicitly stating in text

  The traditional 'See' reference for redirection is a specimen.

- Using controlled vocabularies

  Thesaurus uses the code 'USE' to redirect from uncontrolled to controlled search keys

- Data modeling (relational databases)

  In relational database environment, tables (for Names, Titles, Imprints, etc.) are linked with Record IDs or Pointers, as structured in data model.

- Sharing metadata (identifiers)

  Dublin Core (DC) Metadata serve as links to other web pages, say, generated by same creator, same publisher, or other DC IDs.

- Linking in hypertext

  The Universal Resource Locator (URL), Universal Resource Identifier (URI) are used universally as connecting links between HTML pages, between parts within a HTML page.

In the context of digital resources, the basis for this linking is the identification of information that can be extracted and used to search and locate information within a KOS. This being quite an involved area of discussion, we need a more sophisticated definition of linking to avoid ambiguity and confusion, and admit the following one as a better alternative:

- Relation between Journal and Periodicity is a natural relationship.
- *made up by us*

Relation between an uncontrolled keyword, 'Bharat', to controlled keyword 'India' is a made up relation.

- *deduced from statistics*

Total Number of defaulters reported just-in-time based on circulation statistics.

12.4.2 Types of Linking

Reference Linking

The reference linking is the class of links that can be described as *linking* from metadata (reference, citation) to the full-content.

Some common examples of reference links:

- From an A&I database record to the full-text
- From a citation included in a document to the full-text
- From an OPAC record to an e-journal TOC with further linking possibilities.

Reference links usually target one specific copy of the full-content entity. But the user might rather need or prefer:

- Full content from another supplier
- An OPAC holdings description
- A copy ordering / ILL service
- Another metadata description / abstract
- A book review or access to a net bookshop
- A "full web" search.

Expanding Codes to Full-Text

Coding schemes facilitate communication within a defined group of specialist members. Every discipline has one or more coding schemes to help them communicate precisely and economically. A KOS may use links to connect these coding schemes to the full names for which the code stands. This is an example of static link as opposed to dynamic link that works on-the-fly or just-in-time principle.

The examples provided here include links between databank registration codes and the biological sequence data, and between industrial codes and the full name that the code represents.

Linking to Descriptive Records

Entity names, such as personal and corporate names, location, etc. are linked to additional information about that entity. This was one of the first uses of hyper-linking. KOSs such as dictionaries, glossaries, and classification schemes can be used 'to link the entities in one resource to richer descriptions of that entity in another resource'.

Linking Personal Names to Biographical Information

A common type of authority file is the personal name authority, which controls variants of personal names. For example, the Library of Congress Name Authority File (LCNAF) is used to control variant personal names for authors, editors, artists, and others. The Union List of Artist Names (ULAN), developed by the Getty Vocabulary Program, is another example. Name authorities serve as tools for catalogers and indexers. They ensure use of proper form of name and bringing together all works by or about the person.

A name authority file can also be used to link a bibliographic record or document containing the person's name to a variety of other related materials. If the digital library's resource has a standardised form of the name, it can be identified and searched against the authority file to locate variants. The standardised and variant forms can be joined in a search against a variety of other resources that can provide related information.

For example, in the case of a digital library of images of artists' works, biographical or critical text, a name authority file such as the ULAN or the LCNAF can act as an intermediate file to provide additional information.

Linking Individual Industrial Codes to the Full Scheme

The SIC codes have been used by the U.S. government, economists, financial markets, regulators, and procurement offices to identify manufacturing, agriculture, and service sectors of the economy. The digital library can provide related information by using the authority files for the coding schemes as *a linked authority file*. If a company or economic sector mentioned in the digital library's collection can be linked to an SIC or NAICS code, the code can be searched against the official tables of definitions maintained by the U.S. Census Bureau. These files provide definitions of the codes and place each code in the classification scheme with other economic sectors.

The digital library's content can be further enhanced by making a link between the SIC and NAICS codes. If the digital library resource has the SIC code, it can be extracted and searched against the Census Bureau's *1997 NAICS and 1987 SIC Correspondence Tables.* The table returns the corresponding code from the alternate scheme.

Linking Organism Names to Taxonomic Records

Genus-species names are the Latin names for organisms e.g., plants, animals, and microorganisms. In taxonomy, living organisms are studied and classified. Records are created for each of these organisms. Generally, these records are linked relationally to the other organisms in a hierarchy. Beyond the organism name and the information that it and its placement in the hierarchy convey, taxonomic records use other elements to describe the organism.

These may include distribution patterns, the authority for naming and classification, and the date the organism was identified. Scientists base the information on specimens that are retained because they serve as the physical evidence of the description. Natural history museums, private collections, and individual scientists assign number, or codify the specimens in their collections.

Linking Sequence Numbers to Bio-sequence Databanks

National Center for Biotechnology Information is the most frequently used referenced databanks on the Web. They include GenBank and the Research Collaborators for Structural Bioinformatics Protein Data Bank. Each sequence number is different, but all begin with a persistent code identifying the databank. The link between the literature and the databank is made the following ways. Through a search profile, a text analysis program, or keyword indexing, the text is analysed and the sequence databank numbers identified.

An active link consists of a search strategy can be embedded to locate that sequence number in the databank where the actual sequence is stored. When the user clicks on the active link, the script is generated and launched from the user's browser. The Web-enabled database is searched, and the sequence record is returned to the user.

This type of connection exists between the National Library of Medicine's (NLM) search service, PubMed, and GenBank. If a PubMed search hits records that bear GenBank numbers, an automatic search on GenBank is triggered resulting display of the sequence records.

Linking Chemical Names to Molecular Structures

There are competing systems of nomenclature (i.e., that of the Chemical Abstracts Service [CAS] and of the International Union of Pure and Applied Chemistry) as well as common and commercial synonyms.

BIOSIS, the world's largest not-for-profit producer of biological and biomedical databases uses the chemical registry number to link chemical names with molecular structures. In 1993, BIOSIS began processing its bibliographic citations (titles and keywords) to automatically identify chemical names. BIOSIS assigns CAS Registry Numbers (RNs) to the chemical names identified in this process.

Linking Entity Names to Physical Specimens

Sometimes, we may require going beyond linking the related digital resources, and connect entity names in the digital library resources to physical specimens. Exhibition catalogues describe the art exhibits. Museum catalogues describe objects of art, natural history specimens, and cultural objects. When converted into a computerised database, applications of KOSs become critically important for retrieving information about related objects, and for locating the physical objects.

In natural history community, efforts have been made extensively to create and organise databases of photographs of specimens. The records in their database include the Object Identifiers to facilitate retrieval. The publication of identification codes in the journal literature is also changing. The level of specificity of the identification code changes depending on the biological discipline it belongs to. Vertebrate journals provide the code to the specimen level, Botanical journals tend to list only the institution and the catalogue.

To bring about unified global networked resources for providing unconstrained access worldwide, the following linking strategies are recommended:

1) Use persistent identifiers

2) Use open linking architectures

3) Implement extended services

Self Check Exercises

Note: 1) Write your answers in the space given below.

2) Check your answers with the answers given at the end of this Unit.

7) A link is an expression of relation. Describe the ways links express relationship between two elements. In context of digital resources, how do you define a link?

....

....

....

....

8) Why linking to descriptive records is necessary? How Personal names are linked to Biographical Information? Describe with examples.

....

....

....

....

12.5 UNIVERSALACCESS TO HETEROGENEOUS NETWORKED RESOURCES

The Web is the world's largest mass of bits and bytes. It is too big, and going bigger by 1.5 million pages added every day. Here is the *Internet Commons* that embraces many formal and informal Resource Description Communities – the place where disparate communities are to communicate. KOSs are means of connecting heterogeneous resources of these disparate communities.

"Web-based access to digitised images and their descriptions, at anytime from anywhere, lowers the barriers for access to information resources." (Antoine Isaac, 2007)

Over many generations, librarians, curators and archivists have developed controlled vocabularies such as thesauri, classification schemes and ontologies. These Knowledge Organisation Systems (KOS) take 'a term-based approach, where terms from natural language are the first-order elements of a KOS'. Therefore, KOS works well only when the semantics and syntactic structure of the terms are known to the users, which is very unlikely condition. By merging collections without taking care of the semantic heterogeneity, KOSs shift the burden of search to users to obtain their desired objects from variant collections.

KOSs can be used to (1) provide *alternate subject access*, (2) *add modes of understanding* to digital library resources, (3) support *multilingual access*, and (4) supply terms for *expansion of free-text* searches in domains that are relatively unknown to the user.

12.5.1 Alternate Subject Access

Alternate subject access refers to the provision of additional subject orientations that make the resources accessible to different audiences. Instead of using one single conceptual vocabulary for querying or browsing the objects of both collections simultaneously, users are expected and required to use the terminology of the first KOS to identify objects of the first collection, and the second KOS to identify those of the second collection.

This approach is particularly valuable when the digital library resources appeal to groups that do not share a common terminology. It can be a system of subject headings, a classification scheme, or any other subject-oriented system. Alternate subject access can be provided by

- indexing or classifying the resources using multiple schemes,
- retaining original schemes from organisations that contribute to the digital library, or
- mapping between the primary scheme and an alternate scheme.

These Alternate KOSs are not interoperable at the semantic level. To enhance the interoperability we need to solve heterogeneity problems of two types:

- **Representational heterogeneity**: Vocabularies may be presented in different formats; for example, in XML and in plain text. The guiding models may not be compatible. Their general information needs (e.g. 'terms' in thesauri, 'classes' in classification schemes), and KOS may use different kinds of labels for identifying conceptual entities.

- **Conceptual heterogeneity**: In two different vocabularies we find similar concepts having identical meanings attached with different labels or names. (e.g. like "Virgin Mary" and "Madonna"). Also, there will be concepts that are more general than others (e.g. like "Mother" and "Virgin Mary"). By determining the relations between the concepts and their variant labels, or naming, an integrated system can provide users with seamless access to the content described by several vocabularies.

Cataloguing/ Indexing with Multiple Schemes

Classifying and cataloguing the resources with multiple schemes is the most direct method for providing alternate subject access to a collection. The method, however, is expensive, as it involves employment of cataloguers knowledgeable in relative schemes, modifications to the cataloguing tools and procedures, and far more processing time. If, however, the task of cataloguing does not involve cataloguing at the level of individual books but of individual classes of books requiring modifications of labels only, this method may be found quite acceptable.

Indexing from Contributors

The resources of a digital library are built up commonly with contributions from external specialty sources. These sources often follow KOSs developed by themselves for organising their collections. For example, the IEEE Computer Society uses a proprietary classification system, which can be borrowed for using as *alternate system* for classifying resources on computer science. NASA database on aeronautics and astronautics includes relevant bibliographic records from U.S. Department of Defense and US Department of Energy, who permit NASA to use their controlled vocabulary terms to create candidate indexing terms for review by NASA's indexers. The terms collected from other organisations can be viewed as an alternate access point, so that at least part of the collection is accessible through another discipline's terminology.

Mapping Multiple Schemes

Mapping one or more schemes is an indirect method for providing alternate subject access. As reported by Gale Hodge, the experience of BIOSIS, the world's largest private sector abstracting and indexing service in life sciences, illustrated best an application of this method. The records that BIOSIS contributes to TOXLINE database of NLM are processed automatically to have appropriate terms added. This is based on a mapping of the natural language terms that occur in the toxicology literature. BIOSIS' normalised natural language keyword indexing with the MeSH terminology. In the new BIOSIS relational indexing structure, BIOSIS builds and maintains authority files that connect natural language disease names to the MeSH-controlled disease terms. When the BIOSIS indexer assigns the free text keyword for the disease name, the appropriate MeSH term is also added to the record as an alternate access point. (BIOSIS 1999). The assignment is based on the development over time of a mapping between the terminology used by BIOSIS and the MeSH-controlled terms.

In addition to providing alternate access points to BIOSIS products, the inclusion of the MeSH terms makes it possible to perform cross database searching on the indexing field with MEDLINE and other databases that include MeSH terms. The inclusion of terms from an alternate KOS, such as MeSH, therefore supports the use of BIOSIS by medical librarians and practitioners who are familiar with MeSH terminology. Unified Medical Language System (UMLS) is a meta-thesaurus developed by the NLM. It is more extensive that links more than 40 separate KOSs from various medical specialties.

Adding New Modes of Understanding to the Digital Library

Many digital library projects remain text-based, or text-as-image-based, some audio, some video, and increasingly more multimedia-dominated as Internet's capability of presenting information in a variety of other modes goes high up in leaps and bound. KOSs can be used to deal with these new dimensions. In the digital library environment, these dimensions can be viewed as layers that can be added on top of one or more objects. Various tools and services can be developed that are geared to a particular mode. For example, the results of a text search can be presented in graphical or visual form for best satisfaction of the users.

A 'geolibrary', which is defined as a digital library holding 'geoinformation', needs a *geospatial dimension* to be added to provide access by place, called *georeferencing*.

Disciplines like ecology, environmental science, political economy, public health and epidemiology, should be greatly benefited by using KOSs that retrieve geoinformation.a digital library with access to such a digital gazetteer service. Through a geospatial KOS, users can see connections between disparate data, because the data are presented in an alternate mode.

Accessing Multilingual Resources

The problems of accessing multilingual text-based resources are of two kinds, the script and the vocabulary. Until very recently, for writing and reading non-English texts, the knowledge workers had to depend on Romanisation with applications of diacritical marks. It is no more necessary today when we can painlessly write and read, using Unicode, any script of the world in its original form, and sort them all in lexical order.

The problems related to the vocabularies are by far more intricate because of the cultural bearings on their semantic and syntactic properties.

KOSs can support the use of digital libraries by disparate communities providing multilingual access. A variety of sources, including multilingual dictionaries and multilingual thesauri, can support this type of access. One of the most extensive multilingual thesaurus efforts is the Generalised Multilingual Environmental Thesaurus (GEMET) from the European Environment Agency (EEA), The GEMET is available in 12 languages, and plans for more to add.

Expanding Free-Text Search Terms

Free-text searching is the most popular method of discovery on the Web, and in view of some experts, it has more potential for versatile applications than controlled-vocabulary-based searching has. This is mainly because, vocabulary control is highly time-consuming, labour-intensive process and it tends to remain outdated at any point of time.

Apart from a relatively small proportion of metadata and controlled vocabulary, the Web is full of natural language elements. However, variations in natural language make free-text searching problematic. The problem becomes more sensitive when search takes place in interdisciplinary areas, or when the user is not quite familiar with the topic. To overcome these terminology differences, KOSs may help selection of free-text keywords.

The Getty Vocabulary Project emphasises support for searching as a significant application of its vocabularies. Harpring (1999) It reports that the vocabularies are increasingly being used in search engines to look for different terms that refer to the same concept. The Getty vocabularies used in the Art and Architecture Thesaurus, the Union List of Artists Names, and the Thesaurus of Geographic Names, are particularly rich in equivalence relationships. Getty developed a prototype called *a.k.a.* to experiment with the use of equivalence terms to broaden or narrow searches across databases on the Web.

KOSs can be very powerful in supporting free-text searching within digital libraries and in integrating Web resources into existing digital libraries. KOSs have generally been developed for a specific discipline, task, or function, or for the indexing of a specific collection or database. Therefore, depending on the domain in which the KOS is being used and the complexity of the system, it may or may not suggest relevant free-text terms. Expanding a search with related terms, rather than pure synonyms, may return hits that are only peripherally relevant to the user. Hodge (2000)

Self Check Exercise

Note: 1) Write your answers in the space given below.

2) Check your answers with the answers given at the end of this Unit.

9) Where is an alternate subject access the most effective approach? How can its interoperability be enhanced?

....

....

....

....

10) What is cataloguing with multiple schemes? How does this method differ from mapping multiple schemes? Describe with illustrations.

....

....

....

....

12.6 FUTURE OF KNOWLEDGE ORGANISATION SYSTEMS ON THE WEB

12.6.1 Semantic Web, Knowledge Organisation, and Conceptualisation of Things

The World Wide Web, or the Web, is an embodiment of human knowledge. Technically speaking, the Web comprises all the resources and users on the Internet that are using the Hypertext Transfer Protocol (HTTP). The Semantic Web is an extension of the current Web designed to help us to find, share, and combine information more easily through the process of Knowledge Organisation (KO).

Classification and KO is about the *grouping of like things*. In LIS, “things” means physical documents or their parts or their representations as well as concepts used to characterise documents. Any grouping of objects is based on conceptualisations of the things. (Birger Hjørland)

From KO perspective, the goal is decided in controlled vocabularies and classifications. For reaching the goal of the information system, a certain concept (or meaning) is selected while other concepts/meanings/views are repressed. Concepts are generally *functional* in facilitating the users' interaction with the natural and cultural environment, and they serve pragmatic purposes. A concept is a way of classifying some part of the world (e.g. plants).

Libraries and museums of the world carry a long cultural heritage, and inherit a distinguished culture of tool making for organising large collections of objects such as books or museum artifacts. These tools are generally referred to as "knowledge organisation systems" (KOS). Different families of KOSs, such as "thesauri", "classification schemes", "subject heading systems", "gazetteers", "lexical databases", "ontologies", and "taxonomies" are widely recognised and applied in both modern and traditional information systems. KOSs attempt to model the underlying semantic structure of a domain. Modern digital information systems afford more options for mapping and presenting alternative orders of information than traditional physical libraries, and offer more possibilities of presenting information from variant points of interests and discourses. "Thus, the challenge is as much intellectual as technical when we want to develop knowledge organisation systems that are useful and meaningful for the end-users operating in complex, interdisciplinary knowledge domains". (Web4lib at ECDL 2006)

12.6.2 Problem Issues Waiting for a Solution

The future of KOS will depend on getting hold of innovative technologies, and collective and collaborated human endeavors to defeat the intricate problems, particularly, focused on the following issues:

- User-centric design strategies for KOS. The question is:
- How to develop understandable and systematic descriptions of concepts and terms?
- How to show and explain relationships? The challenge is to find the appropriate level of explanation, clarity and conciseness.
- How to achieve these in networked situations?
- KOS Interoperability: Cross-browsing and cross-searching between distributed KOS services, mapping between terms, classes and systems, mapping between KOS and ontologies.
- How to achieve semantic interoperability?
- KOS representations and service protocols: A basic infrastructure is needed in order to achieve programmatic access to KOS services. We need to provide protocols for networked access to a variety of vocabularies for different end users and applications. These require standard representations in formats such as RDF/XML.
- What is the appropriate granularity of base services to apply in evolving Web/Grid environments?
- Why and how is the scalable and sustainable management of KOS mappings required?
- Terminology services: We need to identify and specify terminology services for different applications, within a service-oriented approach/architecture, building on the basic infrastructure.

- Social tagging: Participative user-based approaches to knowledge organisation and cataloguing are emerging and attracting significant community support. Social tagging, which is also known as collaborative tagging, social classification, and social indexing, allows ordinary users to assign keywords, or tags, to items. Unlike traditional classification, social tagging keywords are typically freely chosen instead of using a controlled vocabulary. With the development of Web 2.0, where multi-author contribution will be common, collaborative tagging becomes a popular concept. The tagging metadata contributed by users are used for individual-based activities like searching, filtering, navigating, and group-based social networking like finding people with common interest and so on. Tagging data will help to reveal the knowledge sharing and topic networks based on the relations among tagging words.
- What is the role of social tagging and informal knowledge structures versus established KOS?

12.6.3 Semantic Web Activity

The World Wide Web Consortium (W3C), an international consortium is primarily responsible for the creation and development of Web standards and guidelines. With the development of the Resource Description Framework (RDF), Web Ontology Language (OWL), and Simple Knowledge Organisation System (SKOS), the W3C Semantic Web Activity promotes the deployment of technologies for expressing, exchanging and processing metadata in a form processable by machines. The Dublin Core and related vocabularies of the Dublin Core Metadata Initiative (DCMI) represents a crucial contribution to this growing suite of standards.

The W3C's Semantic Web Activity has stimulated a new field of integrative research and technology development, at the boundaries between database systems, formal logic and the World Wide Web. One facet of the Semantic Web vision is the hope of better organising the vast amounts of unstructured (i.e. human-readable) information in the Web, providing new routes to discovering and sharing that information. [SKOS Reference: 2009]

W3C formally defined RDF and OWL as knowledge representation languages; determined their ways of expressing things meaningfully, and provided structures to information already present in the Web. In addition to precise descriptions, application of these technologies over large bodies of information requires the construction of detailed "maps" of particular domains of knowledge.

12.6.4 Simple Knowledge Organisation System

SKOS is a data sharing standard, bridging several different fields of knowledge, technology and practice. SKOS aims to provide a bridge between disperse communities and the Semantic Web, by transferring existing models of knowledge organisation to the Semantic Web technology context.

Looking to the future, SKOS occupies a position between the exploitation and analysis of unstructured information, the informal and socially-mediated organisation of information on a large scale, and the formal representation of knowledge. It is hoped that, by making the accumulated experience and wisdom of knowledge organisation in the library and information sciences accessible, applicable within and transferable to the technological context of the Semantic Web, in a way that is complementary to existing Semantic Web technology, SKOS will enable many new and valuable applications, and will also lead to new integrative lines of research and development in both technology and practice.

The SKOS data model views a knowledge organisation system as a concept scheme comprising a set of concepts. These SKOS concept schemes and SKOS concepts are identified by URIs, enabling anyone to refer to them unambiguously from any context, and making them a part of the World Wide Web.

SKOS concepts can be labeled with any number of lexical (UNICODE) strings, such as "romantic love" or "れんあい ", in any given natural language, such as English or Japanese (written here in hiragana). One of these labels in any given language can be indicated as the "preferred" label for that language, and the others as "alternate" labels. Labels may also be "hidden", which is useful e.g. where a knowledge organisation system is being queried via a text index. Lexical Labels for more on the SKOS lexical labeling properties.

SKOS concepts can be assigned one or more notations, which are lexical codes used to uniquely identify the concept within the scope of a given concept scheme. While URIs are the preferred means of identifying SKOS concepts within computer systems, notations provide a bridge to other systems of identification already in use such as classification codes used in library catalogs.

12.6.5 Future

Future must corroborate deployment of Semantic Web methods in support of Knowledge Organisation systems and services. The assumption underlying semantic digital libraries is that full-text search cannot be the entire solution for the massively expanding information space of emerging digital libraries. Next-generation digital library systems must also be able to handle well-defined metadata describing the stored contents and provide machine support for the end users in their search for content. One crucial first step is to organise bibliographic metadata for automated interpretation by machines. This future perspective of KOS was deliberated in Web4lib at ECDL 2006.

Self Check Exercise

Note: 1) Write your answers in the space given below.

2) Check your answers with the answers given at the end of this Unit.

11) What is W3C? Briefly describe the Semantic Web Activities of W3C.

....

....

....

....

12) Who invented SKOS? What it is and what it is supposed to do? Describe SKOS concepts.

....

....

....

....

12.7 SUMMARY

The theme of this Unit is 'conceptual changes in LIS approaches toward knowledge organisation, as an impact of technology'. We discussed six different topics touching upon different aspects of Knowledge Organisation Systems (KOS).

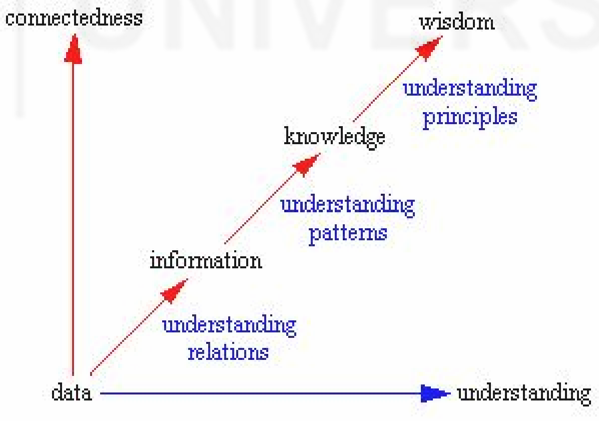
Origin and types of KOS, Analysis and planning of KOS, Methods of linking dispersed digital resources, Methods of accessing heterogeneous networked resources, Contributions of W3C in developing standards like Resource Description Framework (RDF), Web Ontology Language (OWL), and particularly, Simple Knowledge Organisation Systems (SKOS), were highlighted. Lastly, the future of semantic activities is discussed in context of some identified problems yet to be resolved.

12.8 ANSWERS TO SELF CHECK EXERCISES

1) *Information* is one of the most frequently used words in our society, and widely differs in its meaning – from news to intelligence, data to knowledge. However, in context of library and information science, information is understood as something more than data. Data is unrelated, information related.

   Information reflects an organisation of data. Characteristically, it provides verifiable statements of facts that can be either true or false. Knowledge, on the other hand, belongs to a higher plain of organisation, assumes a verified status of truth or falsity, and comprehends generalized pattern of information.

   The following diagram represents the transitions from data, to information, to knowledge, and ultimately to wisdom. Transition from a lower stage to higher up is carried on through 'understanding', which has no separate level of its own. The structural and functional relationships between data, information, knowledge, and wisdom are represented in the following Data-Information-Knowledge-Wisdom (DIKW) model:



2) We may define Knowledge Organisation Systems, in a general way, as semantic tools that present the organised interpretation of knowledge structures. In this broad sense, libraries, encyclopedias, academic disciplines and such other knowledge organisation systems may serve as examples of KOS.

   Common Characteristics in KOS can be described from two view points as follows:

   KOS imposes a particular view of the world.

- The same entity can be characterised in different ways depending on the KOS that is used.
- There must be a sufficient commonality between the concept in KOS and the real world objects it refers.
- A person seeking relevant material by using a KOS must be able to connect his or her concept with its representation in the system.
- KOS imposes a particular view of the world on a particular collection through.
- Providing a controlled list.
- Controlling synonyms or equivalents.
- Linking DL resources to related resources.
- Making semantic relationships explicit.
- Provide a controlled list.
- The KOS imposes a particular view of the world on a collection and the items in it.
- The same entity can be characterised in different ways, depending on the KOS that is used.
- There must be sufficient commonality between the concept expressed in a KOS and the real-world object to which that concept refers that a knowledgeable person could apply the system with reasonable reliability.

3) By examining the trends in scholarly persuasions under the socio-economic and technological influences, we have traced the stages of KO's development with their characteristic features.

Antiquity: Lists

- Middle Ages: Inventories
- Seventeenth Century: Finding lists
- Nineteenth Century: Collocating devices
- Twentieth Century: Automation and Codification
- Twenty-first Century: Social and Collaborative Tagging

Each represents a turning point that reflects changing societal information demands and the development of new technologies. At the beginning, there were only straight forward list of items, often without any order. During the Middle Ages, making of inventory lists were introduced for recording the stock against serial numbers. Then in the next phase, the location marks were added to the inventory for retrieval. Cataloguing as collocating device was introduced in the 19th century. From the second half of the 20th century information technology has taken the lead, backed by computerised database, the Internet technology, Web resources, and new Metadata tools, in achieving the mission and objectives of Knowledge Organisation. Today, in the 21st century a social dimension has been added to take care of the personal informational requirements of individuals, besides the collective demands of the human societies.

4) Generally, KOS may be categorized into three main groups, namely, Term Lists, Classification & categorisation, and Relation Group. The names of KOSs are enlisted below:

   Term Lists

   Authority Files

   Glossaries

   Gazetteer

   Dictionaries

   Classification and Categorisation

   Subject Headings

   Classification Schemes, Taxonomies

   Relationship Group

   Thesauri

   Semantic Networks

   Ontologies

The Semantic Networks and Ontologies are briefly described here:

Semantic Networks

Semantic network is one of the most significant developments in the area of natural language processing technology. Here the concepts and terms not positioned as hierarchies but as a network or a Web. Concepts are thought of as nodes with various relationships branching out from them. These include specific relationships of whole-part, cause-effect, parent-child, etc, instead of standard BT, NT and RT. Princeton's WorldNet is a grand example of semantic network.

Ontologies

In general, ontology is the study or concern about what kinds of things exist - what entities there are in the universe. In information technology, ontology is the working model of entities and interactions in some particular domain of knowledge or practices, such as electronic commerce or "the activity of planning." We may define Ontology as specification of conceptualisations, used to help programs and humans share knowledge. Ontologies are being developed as specific concept models by the Knowledge Management community.

5) Planning for KOS needs to be done carefully to offset all possible risks of unsatisfactory functioning in closed zone environment.

   In the context of the digital library, the physical location of the KOS is a critical factor for deciding upon the architecture of KOS. It is on the physical location the position of KOS depends – whether it will be held externally or internally. There are pros and cons to either approach. If KOS is held internally, that is in controlled zone, Classification, thesaurus, ontologies need to be hand-picked cautiously by applying selection criteria, professional expertise and users' feedback. We should safeguard against all possible risks:

- Selection might limit the access.

This may happen due to various reasons; e.g., when a less inefficient classification scheme is chosen for shelving arrangement, when class numbers are pre-coordinated leaving little scope for libraries to accommodate new concepts; when a chosen subject heading scheme fails to provide unaffected distribution of concepts, e.g. disproportionate distribution of world religions, or other cultural bias; when primary search terms and their spellings are not guided by international standards; and the like shortcomings of KOSs.

- Selection might be biased.

This may happen when there is no well-documented articulate selection policy to follow; or when standard selection criteria are ignored to give preference to a biased judgment.

- Knowledge organisation supports pre-existing concepts, not for new concepts.
- The traditional semantic tools are inflexible and restricted to the limit of possibilities they have had at starting. It must be ensured that KOSs have in-built mechanism to accommodate new concepts, terminology, and even search features to perform unaffectedly.
- Not "user-oriented" – individual user's needs are different.

This may be looked upon as most unfortunate failure from the view point of prevailing Library 2.0 principles where importance of individual requirements of a user has been duly stressed. Libraries must take cognition of the personal needs as well as the collective needs of their users while planning for KOSs.

6) The ways a KOS presents itself is no less important than KOS content. It is true for the textual and graphical presentation on screen, and for the design of the interactive user interface.

Textual Representation

KOS architecture must accommodate the character sets of the incoming sources. This is particularly important if a data string, in ASCII or mark-up language, has been used to represent special characters and diacritical marks. Systems that have been developed in Unicode, which generously accommodate all the existing scripts of languages worldwide have decisively removed the age-old constraints of inter-lingual communications. With the support of Unicode standards, all electronic and web-based KOS platforms are now empowered to communicate directly in original language scripts, that is, without resorting to diacritical marks – a clumsy means of phonetic representations based on extended ASCII. Although the Unicode is now universally accepted standard in the software industry, we still find instances of exceptions. Therefore, the Unicode compliancy needs to be ensured even today when going for a new KOS.

Visibility and Presentation

From the user point of view, the ways a KOS presents itself is no less important than its content is. Because, an ill-designed view front fails to convey the content meaningfully, Moreover, a poor presentation may discourage a person to use the KOS. Therefore, while deciding which KOS should be used and what functions it should serve, the digital library will need to determine how to present the KOS to its users.

The KOS can be exposed to the user in different ways. In the website of the digital library, KOS-related themes or categories may be clustered logically for users' view and interactions. The KOS may be used at a higher level to identify specific portals launched by different user communities. If the content of the digital library includes metadata records, the KOS may be displayed as index terms on the records or as an independent navigation tool. KOS may also remain transparent. For example, a database search procedure may employ a thesaurus behind-the-scene to dig up synonyms for using as multiple search keys. This way, KOSs make presentations of real time statistical information keeping the calculation system behind the scene.

7) There are many ways of expressing relation. Not all are hypertext links, and some follow quite different methods, such as:

- Citing together
- An article cites more than one item; the cited items are related with each other in reference to the citing source.
- Explicitly stating in text.
- The traditional 'See' reference for redirection is a specimen.
- Using controlled vocabularies.
- Thesaurus uses the code 'USE' to redirect from uncontrolled to controlled search keys.
- Data modeling (relational databases).
- In RDB environment, tables (for Names, Titles, Imprints, etc.) are linked with Record IDs or Pointers, as structured in data model.
- Sharing metadata (identifiers).
- Dublin Core (DC) Metadata serve as links to other web pages, say, generated by same creator, same publisher, or other DC IDs.
- Linking in hypertext.
- The Universal Resource Locator (URL), Universal Resource Identifier (URI) are used universally as connecting links between HTML pages, between parts within a HTML page.

In context of digital resources, the basis for this linking is the identification of information that can be extracted and used to search and locate information within a KOS. This being quite an involved area of discussion, we need a more sophisticated definition of linking to avoid ambiguity and confusion, and admit the following one as a better alternative:

A link is a connection from one page to another destination such as another page, or a different location on the same page

8) Linking to Descriptive Records is necessary as because pieces of information related to any particular target item, or entity, are usually stored at different locations, as a part of other records, and those are required to be retrieved by the search system. Entity names, such as personal and corporate names, location, etc. are linked to additional information about that entity. This was one of the first uses of

hyper-linking. KOSs such as dictionaries, glossaries, and classification schemes can be used 'to link the entities in one resource to richer descriptions of that entity in another resource'.

Linking Personal Names to Biographical Information

A common type of authority file is the personal name authority, which controls variants of personal names. For example, the Library of Congress Name Authority File (LCNAF) is used to control variant personal names for authors, editors, artists, and others. The Union List of Artist Names (ULAN), developed by the Getty Vocabulary Program, is another example. Name authorities serve as tools for catalogers and indexers. They ensure use of proper form of name and bringing together all works by or about the person.

A name authority file can also be used to link a bibliographic record or document containing the person's name to a variety of other related materials. If the digital library's resource has a standardised form of the name, it can be identified and searched against the authority file to locate variants. The standardised and variant forms can be joined in a search against a variety of other resources that can provide related information.

For example, in the case of a digital library of images of artists' works, biographical or critical text, a name authority file such as the ULAN or the LCNAF can act as an intermediate file to provide additional information.

9) *Alternate subject access* refers to the provision of additional subject orientations that make the resources accessible to different audiences. This approach is particularly valuable when the digital library resources appeal to groups that do not share a common terminology. It can be a system of subject headings, a classification scheme, or any other subject-oriented system. Alternate subject access can be provided by

- indexing or classifying the resources using multiple schemes,
- retaining original schemes from organisations that contribute to the digital library, or
- mapping between the primary scheme and an alternate scheme
- These Alternate KOSs are not interoperable at the semantic level. To enhance the interoperability we need to solve heterogeneity problems of two types:
- **Representational heterogeneity**: Vocabularies may be presenred in different formats; for example, in XML and in plain text. The guiding models may not be compatible. Their general information needs (e.g. 'terms' in thesauri, 'classes' in classification schemes), and KOS may use different kinds of labels for identifying conceptual entities.
- **Conceptual heterogeneity**: In two different vocabularies we find similar concepts having identical meanings attached with differenr labels or names. (e.g. like "Virgin Mary" and "Madonna"). Also, there will be concepts that are more general than others (e.g. like "Mother" and "Virgin Mary"). By determining the relations between the concepts and their variant labels, or namings, an integrated system can provide users with seamless access to the content described by several vocabularies.

10) **Cataloguing with Multiple Schemes**

Cataloguing the resources with multiple schemes is a direct method for providing alternate subject access to a collection. The method, however, is expensive, as it involves employment of cataloguers knowledgeable in relative schemes, modifications to the cataloging tools and procedures, and far more processing time. If, however, the task of cataloguing does not involve cataloguing at the level of individual books but of individual classes of books requiring modifications of labels only, this method may be found quite acceptable.

Mapping Multiple Schemes

Mapping one or more schemes is an indirect method for providing alternate subject access. As reported by Gale Hodge, the experience of BIOSIS, the world's largest private sector A&I service in life sciences, illustrated best an application of this method . The records that BIOSIS contributes to TOXLINE database of NLM are processed automatically to have appropriate terms added. This is based on a mapping of the natural language terms that occur in the toxicology literature. BIOSIS' normalized natural language keyword indexing with the MeSH terminology. In the new BIOSIS relational indexing structure, BIOSIS builds and maintains authority files that connect natural language disease names to the MeSH-controlled disease terms. When the BIOSIS indexer assigns the free text keyword for the disease name, the appropriate MeSH term is also added to the record as an alternate access point. (BIOSIS 1999). The assignment is based on the development over time of a mapping between the terminology used by BIOSIS and the MeSH-controlled terms.

In addition to providing alternate access points to BIOSIS products, the inclusion of the MeSH terms makes it possible to perform cross database searching on the indexing field with MEDLINE and other databases that include MeSH terms. The inclusion of terms from an alternate KOS, such as MeSH, therefore supports the use of BIOSIS by medical librarians and practitioners who are familiar with MeSH terminology. Unified Medical Language System (UMLS) is a meta-thesaurus developed by the NLM. It is more extensive that links more than 40 separate KOSs from various medical specialties.

11) **Semantic Web Activity**

The World Wide Web Consortium (W3C), an international consortium is primarily responsible for the creation and development of Web standards and guidelines. With the introduction of the Resource Description Framework (RDF), Web Ontology Language (OWL), and Simple Knowledge Organisation System (SKOS), the W3C Semantic Web Activity promotes the deployment of technologies for expressing, exchanging and dealing out metadata in a form that machines can process. The Dublin Core and related vocabularies of the Dublin Core Metadata Initiative (DCMI) represents a crucial contribution to this growing suite of standards.

The W3C's Semantic Web Activity has stimulated a new field of integrative research and technology development, at the boundaries between database systems, formal logic and the World Wide Web. One facet of the Semantic Web vision is the hope of better organising the vast amounts of unstructured (i.e. human-readable) information in the Web, providing new routes to discovering and sharing that information.

W3C formally defined RDF and OWL as knowledge representation languages; determined their ways of expressing things meaningfully, and provided structures to information already present in the Web. In addition to precise descriptions, application of these technologies over large bodies of information requires the construction of detailed “maps” of particular domains of knowledge.

12) **Simple Knowledge Organisation System**

Simple Knowledge Organisation Systems (SKOS) is a data sharing standard, created and developed by W3C. SKOS aims to bridge several different fields of knowledge, technology and practice as well as between disperse communities and the Semantic Web, by transferring existing models of knowledge organisation to the Semantic Web technology context.

Looking to the future, SKOS occupies a position between the exploitation and analysis of unstructured information, the informal and socially-mediated organisation of information on a large scale, and the formal representation of knowledge. It is hoped that, by making the accumulated experience and wisdom of knowledge organisation in the library and information sciences accessible, applicable within and transferable to the technological context of the Semantic Web, in a way that is complementary to existing Semantic Web technology, SKOS will enable many new and valuable applications, and will also lead to new integrative lines of research and development in both technology and practice.

The SKOS data model views a knowledge organisation system as a concept scheme comprising a set of concepts. These SKOS concept schemes and SKOS concepts are identified by URIs, enabling anyone to refer to them unambiguously from any context, and making them a part of the World Wide Web.

SKOS concepts can be labeled with any number of lexical (UNICODE) strings, such as “romantic love” or “रोमांटिक प्रेम”, in any given natural language, such as English or Hindi (written here in Nagari). One of these labels in any given language can be indicated as the “preferred” label for that language, and the others as “alternate” labels. Labels may also be “hidden”, which is useful e.g. where a knowledge organisation system is being queried via a text index. Lexical Labels for more on the SKOS lexical labeling properties.

SKOS concepts can be assigned one or more notations, which are lexical codes used to uniquely identify the concept within the scope of a given concept scheme. While URIs are the preferred means of identifying SKOS concepts within computer systems, notations provide a bridge to other systems of identification already in use such as classification codes used in library catalogs.

12.9 KEYWORDS

Classification : The operation of grouping elements and establishing relationships between them (or the product of that operation).

Controlled Vocabulary : A collection of preferred terms that are used to assist in more precise retrieval of content. Controlled vocabulary terms can be used for categorising content, building labeling systems, and creating style guides and database schema. One type of a controlled vocabulary is taxonomy

Data Model : A description of data that consists of all entities represented in a data structure or database and the relationships that exist among them. It is more concrete than an ontology but more abstract than a database dictionary (the physical representation).

Domain : A sphere of knowledge, influence, or activity.

Element : An object or concept.

Extensible Markup Language (XML) : A W3C standard markup language for documents containing structured information. As opposed to HTML, which is designed specifically for web browsers, XML is designed for much wider use and is extensible to fit each application. XML is the basis for an incredible array of standards that describe everything from messages between systems to security specifications to document structures. The advantage of XML is that it is human understandable and platform independent.

Indexing : The intellectual analysis of the subject matter of a document to identify the concepts represented in the document and the allocation of descriptors to allow these concepts to be retrieved. Indexing a large number of documents can be done semi-automatically using software applications.

Metadata : A definition or description of data. In data processing, metadata is definitional data that provides information about, or documentation of, other data managed within an application or environment.

Ontology : Ontologies resemble faceted taxonomies but use richer semantic relationships among terms and attributes, as well as strict rules about how to specify terms and relationships. Because ontologies do more than just control a vocabulary, they are thought of as knowledge representation. The oft-quoted definition of ontology is "the specification of one's conceptualisation of a knowledge domain."

Relationships : A defined linkage between two elements.

Resource Description : A W3C standard XML framework for describing.

Framework Framework (RDF) and interchanging metadata. The simple format of resources, properties, and statements allows RDF to describe robust metadata, such as ontological structures. As opposed to Topic Maps, RDF is more decentralized because the XML is usually stored along with the resources.

Semantic Web : The Semantic Web is an extension of the current Web that will allow you to find, share, and combine information more easily. It relies on machine-readable information and metadata expressed in RDF.

Taxonomy : A classification of elements within a domain

Thesaurus : A taxonomy that also includes associated and related terms. It is the most complex type of controlled vocabulary, and is sometimes used to standardise an organisationís terminology and subsequently inform both navigation and search systems.

12.10 REFERENCES AND FURTHER READING

Bean, Carol A. *Relationships in the Organisation of Knowledge.* Ed. Rebecca Green and A. Bean. New York: Springer, 2001.19-35. Print.

Bliss, Henry Evelyn. *The Organisation of Knowledge and the System of the Sciencesý*. New York: H. Holt and Company, 1929. Print.

Fugmann, Robert, ed. "Tools for Knowledge Organisation and the Human Interface: Proceedings". *INDEKS, 1990. 1st International ISKO Conference*, Darmstadt, 14-17 August 1990. Print.

Gilchrist, Alan, ed. *From Classification to "Knowledge Organisation": Dorking Revisited or "Past is Prelude*". The Hague: FID, 1997. Print.

Hodge, Gail M. *Systems of Knowledge Organisation for Digital Libraries: beyond Traditional Authority Files*. Washington DC: Council on Library and Information Resources. Web 24 September 2012. < http://www.clir.org/pubs/reports/pub91/contents.html>

Hjørland, Birger. "Theory of Knowledge Organisation and the Feasibility of Universal Solutions". Presented at the *Eighth International ISKO Conference*, London, July 13-16, 2004. ISKO, 2004. Print.

Husby, Ole. *Linking Resources.* BIBSYS, 2001. Web. 24 September 2012. < http://old.stk.cz/elag2001/Papers/OleHusby/Show.html>

Joachim, Martin D, ed. *Languages of the World: Cataloging Issues and Problems.* New York: Haworth Press, 1994. Print.

Knowledge Organisation and the Global Information Society. Ed. Ia McIlwaine. Wurzburg: Ergon-Verlag, 2005. Print.

Knowledge Organisation and Change. Ed. Rebecca Green. Frankfurt: Indeks Verlag, 1996. Print.

López-Huertas, Mariá J., ed. *Challenges in Knowledge Representation and Organisation for the 21st Century.* Wurzburg: Ergon-Verlag, 2002. Print.

Rowley J.E. and Richard J. Hartley. *Organising Knowledge: An Introduction to Managing Access to Information.* Aldershot: Ashgate, 2008. Print.

Williamson, Nancy Joyce and Clare Beghtol eds. *Knowledge Organisation and Classification in International Information Retrieval.* New York: Haworth Press, 2004. Print.

UNIT 13 ONLINE CATALOGUES: DESIGN AND SERVICES

Structure

13.0 OBJECTIVES

After reading this Unit, you will be able to:

- define an online library catalogue;
- explain its functions and principles;
- discuss changes in its working; and
- design an online library catalogue.

13.1 PHYSICAL CATALOGUE TO OPAC: CHANGING PERSPECTIVES

13.1.1 Physical Catalogue

Around 4000 years ago, libraries were there in Sumer and other ancient regions. Those libraries stored collections of stone tablets, and had their catalogues as well. Early catalogues were mostly inventory lists, all hand written until libraries stared printing their catalogues in book form. The popularity of printed catalogue grew for their easy handling and storage. However, the librarians experienced difficulties in keeping their catalogues up-to-date. Production of printed catalogue was time-consuming, labour-intensive, and expensive.

Librarians found solution in card catalogue. In 1860, Harvard Librarian John Langdon Sibley introduced a card catalogue for public use. Card catalogue stayed for a century as the most effective form of catalogue.

In early 1960s, Library of Congress developed a system of machine-readable catalogue with a new set of rules for machine manipulation. They did it for transferring their massive bibliographic records into electronic format. Although, the development of online catalogue was not their immediate objective, this initiative had paved way to build online catalogue with support of advancing telecommunications computing technology.

It took more than three decades for libraries for reaching the era of online catalogue.

13.1.2 Descriptive Catalogue

Along with the changes in the physical shapes and media, the rules governing the arrangement of bibliographic descriptive elements have been continuously evolving. The First Age of Descriptive Cataloguing began with Panizzi and included Cutter. Michael Gorman labelled it as “the age of the single-author code.”

Before, Anthony Panizzi’s *Rules for the Compilation of the Catalogue*, printed in 1841, there had been no serious and successful attempt to construct comprehensive code for cataloguing. Many contemporaries of Panizzi like Thomas Carlyle thought of a catalogue as simple listing of authors and book titles. They found his rules unnecessary. The criticism could not prevent him publishing his rules. He thought:

“A reader may know the *work* he requires; but he cannot be expected to know all the peculiarities of different *editions,* and this information he has a right to expect from the catalogues.” So here we have two individuals looking at the same object—the book—but seeing different things.

Charles Ammi Cutter his *Rules for a Dictionary Catalogue*, published in 1876, has deeper impact on the development of cataloguing principles and methodologies. We can trace FRBR's development in Cutter's cataloguing code.

Second Age of descriptive cataloguing was "the era of the committee code", as Gorman called. The year 1908 was important in cataloguing: the American Library Association and the Library Association of the United Kingdom published a set of common cataloguing rules. They did not agree on absolutely everything, so separate American and British editions were made, but this was the first set of Anglo-American cataloguing rules.

Third Age of Descriptive cataloguing began in 1967 with the release of A*nglo-American Cataloguing Rules*. It was an international (American, Australian, British, Canadian) standardisation of descriptive cataloguing rules, with a philosophy based on the *Paris Principles*. The *Statement of Principles* passed at the International Conference on Cataloguing Principles held in Paris in 1961. These principles eventually generated a large set of detailed rules into *Anglo-American Cataloguing Rules* (*AACR*) and serve as the basis of other national codes. It was a culmination of most of the major ideas: *an axiomatic approach*, *user needs*, and *standardisation* and *internationalisation*.

Third Age of Descriptive Cataloguing has been continuing *in spirit* till today. In terms of its applicability of its regulatory principles, its future looks little uncertain, particularly in approaching environment of Web 2.0 technology.

13.1.3 Standards

The three cataloguing standards produced during the Third Age, namely MARC, ISBD and AACR2, are presently being critically reviewed by the professional bodies and experts.

The Machine Readable Cataloguing (MARC) format came out after a long trial, with valid promises to carry on bibliographic record sharing at large scale. The International Federation of Library Associations and Institutions (IFLA) published a standard specification for bibliographic description (ISBD) in 1971. ISBD was originally constructed as a means of standardising the presentation of descriptive data so that it could be machine-translated into MARC.

In 1978, AACR2 reached a historic agreement between the North American and the British library associations and published the two sets of rules as a continuous single text document.

These three standards, which serve as the bedrock of cataloguing as of now, had a beginning in card catalogue environment, and embedded specifications well-suited for the physical catalogue of 5" × 3" cards.

13.1.4 Electronic Catalogue

Michael Gorman considered that the confluence of a need (national and research libraries throughout the world needing less expensive and more current cataloguing) and a means (automation and, more specifically, the MARC format) that has brought us nearer to Universal Bibliographic Control (UBC) than anyone would have dreamt possible thirty years ago.

In early 60s, the Library of Congress had developed a schematic design to help transfer its card catalogue records into newly installed computer. The design provided a way to pass on coded instructions to the computer, made of numbers, alphabets and special characters, along with bibliographic data. The computer requires all those for identifying cataloguing data elements, and reproducing LC catalogue cards, and tapes for distribution among the member institutions. This was how MARC began. The communication format of the MARC records, in the beginning, was suitable for encoding the bibliographic elements of books in roman script. During the span of four decades, the Library of Congress, with collaborators enriched MARC in terms of document-types and contents. They enhanced its structural design to accommodate requirements of all kinds of recorded knowledge– from prehistoric artifacts to Web publications.

For efficient management of bibliographic data in networking environment five separate formats are defined to handle five types of data: bibliographic, holdings, authority, classification, and community information.

From day one of mechanisation, the Library of Congress pursues the policy of sharing benefits of their machine-readable catalogue with libraries having no machine to read, by the way of distributing printed catalogue cards in conventional format. This happened to be the primary reason for structuring MARC in compliance with catalogue card production.

13.1.5 Online Catalogue

Definition of an online catalogue is simple. It is a catalogue in electronic (machine-readable) format accessible online. The term 'online' means 'connected to a computer network or accessible by computer'.

The recent technological boom has empowered libraries to exploit computer and communication facilities to their best advantage, and implement MARC standard for interchange of bibliographic information between databases. Besides institutional will, what the libraries need most is the knowledge and skill to develop MARC compliant databases addressing local as well as global requirements.

As we have already noticed, the formatting of bibliographic elements in a MARC record reflects a semblance of catalogue card, when computer processing systems call for a different approach. Nevertheless, "the fact is that there are tens of millions of MARC records in the world; … MARC is the basis for almost all automated bibliographic systems (including commercially produced systems); and, no practically feasible or demonstrably better system has been advocated."

Cataloguers appreciate these days that their working tools are too many yet inadequate to keeping up with what is happening around them. Today things are vastly more complicated:

- Cataloguing costs money and takes time. Sharing cataloguing records will save both, if everyone can agree on how to catalogue things the same way.
- Electronic resources (on computers) are hard to catalogue and manage, and not always easy to make available.
- Everything comes in many formats, and they are hard to catalogue, manage, and make available, too.
- There is more of everything.
- Technology is changing how libraries work, what they have in their collections, and what users need and expect.

13.1.6 Next-Generation Catalogue

The technology of Web 2.0 has opened up alternative ways to develop many different search models for bibliographic database. Quite a few successful catalogue systems already visible online, though still in experimental stage. They:

- give the patron a simple search interface that allows the user to enter vague, broad, and simple searches;
- allow the patron to drill down through the large result list, narrowing it down by whatever criteria they choose, until it is as precise as they want sort the results list so that the most relevant items are at the top of the list; and
- are tolerant of misspellings and unusual word choices in the patron's search.

The list shows what patrons desire to get, and what the next-generation catalogues can meet. Next-generation catalogues give patrons the same tools they already enjoy on websites. Since library databases are generally built in compliance with MARC, they look beyond integrated library system and will have little problem in working with a variety of systems.

Self Check Exercise

Note: 1) Write your answer in the space given below.

2) Check your answer with the answers given at the end of this Unit.

1) What is Next-generation catalogue? What are their special characteristics.

....

....

....

....

13.2 ONLINE CATALOGUEAND MARC DATABASE

13.2.1 Online Catalogue – Definition

Online catalogue is a library catalogue consisting of a collection of bibliographic records in machine-readable format. In other words, it is a machine-readable catalogue, by definition, and something more. It needs to be maintained in a dedicated computer that provides uninterrupted interactive access via terminals or workstations in direct, continuous communication with the central computer. In short, it is a catalogue in electronic (machine-readable) format accessible online. The term 'online' means 'connected to a computer network or accessible by computer'. For example, when you are connected to the Internet, you are online, when you search British Library Catalogue, you search an *online database*.

Online database needs two *common* record formats, a physical format, and a logical format. The common physical format serves as a vehicle to carry records of bibliographic data to remote computers. The common logical format, on the other hand, ensures that the recorded data are stored and interpreted correctly at remote ends.

13.2.2 What is MARC?

The key to the machine-readability is the common record format. MARC follows a physical structure and a set of control mechanism for enabling the machine to identify and process the data elements, whenever needed. This physical data structuring scheme, originally developed as a carrier of MARC data, is now being used as a standard, nationally (ANSI Z39.2) and internationally (ISO 2709), by other communication formats for bibliographic information interchange, like UNIMARC, CCF, national MARCs, etc. They all are implementations of ANSI Z39.2 / ISO 2709 standard.

The LC MARC found its way to US MARC, as an acknowledged national format for bibliographic communications, and after that evolved into MARC 21. MARC 21 is not a new format but a harmonised edition of USMARC, CAN/MARC, UKMARC and AUSMARC brought out as a consolidated scheme. Although, one finds there little change content wise, MARC 21 differs significantly in its vision. It tends to take issues beyond national preferences, and to negotiate with the up-coming events as well. Being its focus shifted from geographic to temporal zone, MARC 21 appears as a MARC version for the 21st century.

13.2.3 MARC Compliant Database

MARC database and MARC compliant database are not same. A MARC database contains a sequential file of hierarchically arranged records following ANSIZ39.2/ ISO2709 specifications and MARC defined content designation for every field in every record; whereas, a MARC compliant database may be a relational, or a different model, holding records comprising elements equivalent to MARC fields by definition. A database in compliance with MARC can transform its records into MARC communication format and exchange records with any MARC database or with another MARC compliant database; otherwise it cannot.

No online library can function without capability of interchanging records. As things stand now, it is almost impossible for a library to maintain a MARC-incompatible catalogue. Because, the population of records in MARC databases is by far greater than the rest of the databases as a whole. When working as stand-alone electronic catalogues, many libraries feel no obligation to follow a common record structure, overlooking any probability of sharing bibliographic data in future. For time being any non-standard computerised catalogue, system can serve the current and local needs satisfyingly. However, the good and committed libraries are expected to move toward standardisation sooner or later.

13.2.4 Resource Sharing Initiative

In manual environment, libraries had no other options but invest huge staff time in creating catalogue entries of their own for every document they acquire. When MARC files became available, individual libraries, with no computer facility, started buying computer-printed cards from the Library of Congress. In due course, more and more libraries installed computers. Those who designed their databases MARC compliant, preferred to download MARC from the Library of Congress or from other bibliographic utilities, like OCLC, WLN, RLIN, and A-G Canada. The agencies extended their on-line services based on their high-power mainframe systems, against fee-plus-communication-cost. Output was a magnetic tape containing bibliographic information tagged in modified MARC 21 format. The academic institutions, having sufficient fund and sophisticated library software, joined bibliographic utilities. They enjoyed benefit of sharing their collective resources by downloading and contributing MARC records. With the advancement of information technology, computer systems are getting more powerful, affordable, and friendlier day-by-day. Libraries have found automated systems are cost-effective as well as efficient, and resource sharing is the key to solve the problem of unleashed proliferations of information and ever-increasing information price. This happens to be the backdrop of coming of MARC into prominence. The scenario is different in less-advanced countries. Although computer power is within easy reach of the libraries, the concept of sharing MARC records is picking up.

13.2.5 Importance and Advantages

It is important for the implementing agencies to note the plus points of MARC, particularly MARC 21, and its advantages over other bibliographic exchange formats, like UNIMARC, CCF, and National MARCs:

- MARC is a collaborative effort. The MARC scheme, with its massive structure and exceedingly ambitious planning for upgrading its capabilities, depends on active participation of expert agencies. Primarily, the Library of Congress is responsible for creation and up keeping of the scheme. The National Agricultural Library, National Library of Medicine, United States Government Printing Office, and National Library of Canada, together with the Library of Congress serve as sources of authoritative cataloguing and provide most of the codes and standards used for MARC cataloguing.

- MARC is one of the very few international schemes that sustain the tempo of development and continuously update its resources. It stands current and serves as a reliable processing instrument for machine-readable cataloguing. The Library of Congress made no revisions in MARC specifications unilaterally. Two Committees were set up to review and update MARC 21 format documentation. One is the MARC Advisory Committee and the other, Machine- Readable Bibliographic Information (MARBI) Committee of the American Library Association (ALA). The MARC Advisory Committee is composed of the representatives from libraries and scholarly associations, as well as from the bibliographic utilities and vendor groups. The MARC authorities welcome proposals from all professional institutions and experts for incorporating changes in specifications in the interest of the user community at large.

- MARC 21 embraces all media of documents in its scheme. In 1970, the Library issued specifications for magnetic tapes containing *monographic* records in *MARC II format,* as they named it. The first document published with USMARC in title appeared in 1990. After a decade a revised edition of USMARC format for bibliographic data, has been produced in collaboration with the British Library and National Library of Canada as MARC 21. As its variant name suggests the revised format aims to meet the challenge of the 21st century.

- There is hardly anything in MARC that falls outside the scope of standardisation. Being a bibliographic information interchange format, MARC in principle does not dictate use of any particular standard, but allow its users to follow any one of the available standards or authorities, especially for data rendering, choice of access points, codes, classifications and choice of subject heading, name authorities, etc. MARC Program for Cooperative Cataloguing (PCC), as an international cooperative effort emphasises on greater flexibility in tailoring local cataloguing practice to local needs and priorities maintaining relative standards.

- MARC databases are multiplying. The Library of Congress, the official depository of United States publications, is a principal source of cataloguing records for US and international publications. The OCLC shared database, another enormous source of MARC records, available online for downloading against fees. These apart, many a public library and institutional library in USA, Canada, Britain and Australia offer MARC records, often at no cost. The population of MARC databases is growing fast. In India, under Government patronage, MARC 21 implementations are in progress in a number of libraries, including the National Library of India. The increase in accessibility of MARC records tends to boost up MARC implementations in turn.

13.2.6 Schematic Design

In early 60s the Library of Congress had developed a schematic design to help transfer its card catalogue records into newly installed computer. The design provided a way to pass on coded instructions to the computer, made of numbers, alphabets and special characters, along with bibliographic data. The computer requires all those for identifying cataloguing data elements, and reproducing LC catalogue cards, and Tapes for distribution among the member institutions. This was how MARC began. The communication format of the MARC records, in the beginning, was suitable for encoding the bibliographic elements of books in roman script. During the span of four decades, the Library of Congress, with collaborators enriched MARC in terms of document-types and contents. They enhanced its structural design to accommodate requirements of all kinds of recorded knowledge – from prehistoric artifacts to web-publications.

Union catalogues comprising MARC records can be generated and maintained efficiently by using MARC 21 holding format. MARC 21 bibliographic format offers full coverage of bibliographic elements related to every possible type of document. Its integrated design allows records of varied types and sizes to be kept together, and searched together.

13.2.7 Bibliographic and other Related Formats

For efficient management of bibliographic data in networking environment five separate formats are defined to handle five types of data: bibliographic, holdings, authority, classification, and community information.

Bibliographic Format

MARC 21 Format for Bibliographic Data is central to MARC system. It is an integrated format defined for the identification and description of different forms of bibliographic material, no more restricted to monographs as it was, but extends specifications in details for encoding elements of documents of any description, shape and material type — such as books, serials, computer files, maps, music, visual materials, web sites, databases and mixed materials too.

Holding Format

Format for Holdings Data serves as a subset of Bibliographic format. It contains that portion of format specifications which a library may require for encoding data elements pertinent to the holdings and physical location of the items the library holds.

Authority Format

This is a format of the Authority Record. Libraries maintain the files of records as cataloguing tools. Authority records help inputting consistently and uniformly personal and corporate names, subject headings, which serve as preferred access points in records. Format specifications for encoding these bibliographic elements standardise the textual representations of these bibliographic elements and ensure data consistency – a precondition for effective search.

Classification format

MARC 21 Format for Classification Data contains format specifications for encoding classification numbers and their descriptive terms. It serves as an in-house tool for maintaining consistency in library classification, and helps catalogue search by class numbers and browsing books on shelves.

Community Information format

Format for Community Information provides format specifications for records containing house-keeping information, events, programs, services, etc. required for running circulation system and other library management programs.

13.2.8 Implementing MARC

The people working with MARC standards, directly or indirectly, stand responsible for success and failure of MARC in playing its primary role, which is to develop sharable bibliographic resources.

The software developers building MARC utility programs, the database vendors distributing MARC records, and the professionals cataloguing MARC records – must know MARC adequately to do their part responsibly. In implementing MARC compliant library automation system, it is important that the people at top acquire clear-cut understanding about MARC to support their decision-making and help them see that their:

- Library Management Package is demonstratively MARC-compliant, and includes efficient MARC utility programs.
- MARC data files, whenever downloaded from external sources, hold no substandard records.
- MARC cataloguers work proficiently adhering to MARC specifications.

Self Check Exercise

Note: 1) Write your answer in the space given below.

2) Check your answer with the answers given at the end of this Unit.

2) What is MARC? State briefly how MARC21 was evolved.

....

....

....

....

13.3 MACHINE - READABLE CATALOGUING: STRUCTURAL DESIGN

13.3.1 Bibliographic Data Vehicle

Cataloguing data elements are transmitted as contents in an electronic file especially designed as per ISO 2709 specifications for bibliographic information interchange. The file stores a continuous stream of records consecutively placed in a single row, each separated by a special character.

```
record 1▲ record 2▲ record 3▲record 4▲record 5▲ … ▲record n↔
```

Fig. 13.1: Physical File

Computer singles out the records one by one by tracing the predetermined record-separators, which is a special character rarely used in text, often a non-printable sign. Each physical record as specified in ISO 2709 standard comprises three sections of data:

- Leader or Record Label
- Directory
- Data Content

Leader or Record Label

The Leader consists of a string of 24 characters, 00 to 23. The string includes mostly coded information all about the organisation and features of the record itself, only a few about bibliographic information.

```
Position: 01 ⟶―――――――――――――――――――⟸ 24

String:   01218cam  22003134a 45000
```

Fig. 13.2: Leader

The code may be a letter, number, or a Blank Space, as illustrated in Fig. 2. The meaning of any one-lettered code depends on its relative position in the string. For example, letter 'a' in position 06 indicates that the Record is for a language material.

Directory

The Directory serves as road map of Data Contents area. Directory information is dynamically gathered and stored in a place between the Leader and the Data Contents sections. The Directory is used for locating various fields of data elements each represented by a unique address comprising 12 numeric characters made of Field Tag, position, and field length.

```
00100090000000500170000900800410002690600450006792500420011295550
2130015401000170036702000270038402000220041104000180043304200140
0451050002400465082001600489100001900505245003700524260004200561
3000038006034900025006415000021006665040059006875050071007466500
0330081765000120085080000420 0862
```

Fig. 13.3: Directory

Data Content

MARC cataloguing involves

- logical structure of bibliographic record,
- content designation, and
- data content

Before taking up the logical construct of a MARC record, it should be worthwhile to find out what the components of a logical record are, and how those are organised into a machine-readable record serving as functional library catalogue. We discuss content designators first, because that will provide clues for analysis and interpretation of bibliographic records in communication format.

Content Designation

The data units are clustered in Data Content Section. While reading the continuous character string, computer picks up units of data with the help of predetermined control characters, which separate the physical units as fields and subfields. Hence, Fields and Subfields are actually logical names of physical units and sub-units of data.

By the same logic, the actual text, or value, of a data unit is defined as Field Content. The meaning of a data unit, or field content, is established only by putting it in context of Field Tag. The Field Tags, however, are kept in the Directory section, away from their respective data fields, but programmatically linkable. The 3-digit Field Tag, as indicated earlier, represents the type of bibliographic data contained in the field, e.g. Tag '245' signifies that the field contains *title statement* of the record item. This clarifies that the Field Tag serves a logical purpose of recognising the relationship of the field content with the record item. Like Field Tag, there are other tools to analyse and extract the field values.

Physical Control of Data Content

The two control devices, Subfield Delimiter and Field Terminator, work like traffic signals. The Subfield Delimiters (e.g. '$') embedded in field contents alert computer programs to take the immediate next character as a data identifier code *and not* a part of text. This way, the computer continues to read the text string as a subfield till it encounters another Subfield Delimiter or a Field Delimiter.

The control characters used for Subfield Delimiter and Field Delimiter are special characters, most often nonprintable. In MARC 21 records, an upright solid triangle and an upside-down solid triangle serve as the Field Separator and Subfield Delimiter respectively. When a MARC record is printed in Tagged Format the nonprintable characters are transformed into some legible signs by proxy.

```
LDR 01218cam 22003134a 450

001 13020293

005 20030923103827.0

008 021204s2003 nyua b b 001 0 eng

906 __ $a 7 $b cbc $c orignew $d 1 $e ecip $f 20 $g y-gencatlg

925 0_ $a acquire $b 1 shelf copy $x policy default

955 __ $a jb12 2002-12-04 $c jb12 2002-12-04 $d jb04 2002-12-06 $e jb02
2002-12-06 to Children's $a lb00 2002-12-10 $d lb04 2002-12-12 $a aa07 2002-
12-16 $a ps07 2003-08-25 1 copy rec'd., to CIP ver. $f pv06 2003-08-28 CIP
ver to CCD

010 __ $a 2002154957

020 __ $a 0516242946 (lib. bdg.)

020 __ $a 0516278819 (pbk.)

040 __ $a DLC $c DLC $d DLC

042 __ $a pcc $a lcac
```

```
050 00 $a QL737.C23 $b E34 2003
082 00 $a 599.756 $2 21
100 1_ $a Eckart, Edana.
245 10 $a Bengal tiger / $c by Edana Eckart.
260 __ $a New York : $b Children’s Press, $c c2003.
300 __ $a 24 p. : $b col. ill. ; $c 16 x 19 cm.
490 1_ $a Animals of the world
500 __ $a “Welcome books.”
504 __ $a Includes bibliographical references (p. 23) and index.
505 0_ $a Bengal tigers — Cubs — Roaring — New Words — To find out more.
650 _0 $a Tigers $v Juvenile literature.
650 _1 $a Tigers.
800 1_ $a Eckart, Edana. $t Animals of the world.
```

Fig. 13.4: Tagged MARC Record for Visual Representation

Functions of Content Designators

As Content Designators, the Field Tag, Indicators-1 and Indicator-2, and Subfield Code — all contribute to computer performance in reading the content of a bibliographic record meaningfully. The main objectives of the content designation are to support computer programming in -

- Searching and retrieving all identifiable bibliographic data elements.
- Formatting of retrieved data for visual presentation on screen and in print.

Field Tag

MARC reserves a number of Field Tags, or 3-digit codes, each represents a particular type of data. The 3-digit codes are, in fact, abbreviated form of the field names used for describing bibliographic units. Field Tag provides bibliographic format with flexibility. This is an indispensable feature for recording bibliographic elements since many data fields, e.g. Title, Author, etc. *vary in length*. Field Tag, being potentially a repeatable device, supports *multiple occurrences* of any field specified as ‘Repeatable’ in MARC 21 documentation. When a field, e.g. Tag 700, repeats twice it will produce two Alternate Authors fields. The nature of the data content of a field type determines repeatability (R) /non-repeatability (NR) of fields. For example, a bibliographic record supports only one non-repeatable (NR) main entry field 100 for Personal Author.

100 1_ $a Spilsbury, Richard, $d 1963-	NR
020 __ $a 0516242946 (lib. bdg.)	R
020 __ $a 0516278819 (pbk.)	
650 _1 $a Tigers.	R
650 _1 $a Endangered species.	

Fig. 13.5: Examples of Repeatable Field Tags

Indicators

Indicators are two: Indicator 1 and Indicator 2. Both the Indicators work together as Content Designators. Indicator 1 holds the first position at the beginning of a variable data field; the Indicator 2 holds the next position. Each of these provides some supplementary information about the field content, mostly related to visual presentation of the field. Each Indicator holds single-character code. The code may be a numeric, a lowercase alphabetic character, or a blank space. A blank space may mean:

Indicator undefined

- No value provided, or
- A specific meaning assigned.

Here are few examples:

- Value 3 in 1st indicator of field 246 indicates that field content is a parallel title; while value 3 in 2nd indicator indicates that there will be no notes but an added entry under the parallel title.
- A '0' value of 1st indicator in field 245 indicates that the title is the main entry; in field 246 it indicates that the field content is a portion of title.

TAG	INDICATOR	Field CONTENT
246	3	Parallel Title
246	3	No Notes. Added Entry
245	0	Main Entry under Title
245	3	Skip 3 characters in sorting/filing
650	0	LCSH used for Subject Heading

Fig. 13.6 : Changing Values of Indicators in Variant Fields

- A '0' value of 2nd indicator in field 650 indicates that for subject heading LCSH is followed; in field 245 the 2nd indicator value 3 indicates that the title begins with an article with three non-filing characters ('An' and a space). Indicator value 9 is reserved for local implementation.

Subfield Code

Subfield Codes identify data elements within a field for enabling the computer to manipulate each one separately. A Subfield Code is composed of a Subfield Delimiter and a Data Element Identifier. It is a necessary component of Subfield Code, but not a code by itself. A delimiter's function ends with passing a signal to computer predicting the presence of a Data Element Identifier. Data Element Identifier is a code, and has a crucial role to play in analysing bibliographic data content.

Data Element Identifier consists of either a lowercase alphabetic character or numeric character. The character '9' is kept reserved for local use as data element identifier. The order of subfields is specified by the bibliographic standards followed by the cataloguing agency, e.g., cataloguing rules, ISBD, giving priority to MARC 21 specifications wherever available.

245$a	Title statementTitle proper/short title	NR NR
$b	Remainder of title	NR
$f	Designation of vol./issue and/or date	NR
$g	Miscellaneous information	NR
$h	Medium	NR
$i	Display text	NR
$n	Number of part/section of a work	R
$p	Name of part/section of a work	R
$5	Institution to which field applies	NR
$6	Linkage	NR
$8	Field link and sequence number	R

Fig. 13.7: Examples of Subfield Codes with Dollar sign as Subfield Delimiter

Figure 7 illustrates use of numeric character in Subfield Code as processing parameters $6 and $8 besides repeatability (R) and non-repeatability (NR).

The design aspects of MARC, as we examined, cover the physical and logical control of bibliographic elements with the instrumentality of content designators and other control devices. The robust schematic design of MARC as a whole is responsible for the spectacular development of bibliographic network worldwide.

Self Check Exercise

Note: 1) Write your answer in the space given below.

2) Check your answer with the answers given at the end of this Unit.

3) What is Field Tag? What is the use of a Field Tag? Explain with illustration

....

....

....

....

13.4 METADATA TOOLS FOR CATALOGUING NETWORKED RESOURCES

13.4.1 Introduction

Since the time of Panizzi and Cutter, libraries have been evolving a set of tools for gaining bibliographic control of books, journals, and other printed resources. As the volume grows, and formats of resources diversify, principles of organisation undergo change all the time, in order to process their contents and make them accessible. In print-dominated library environment, the principles of organisation were applied manually, and the set of rules they created for 'technical' processing was known as cataloguing. This process is still being used in manual environment. It is effective, but labour-intensive and highly duplicative from library to library, and did not scale well as knowledge expanded. The mechanisation of cataloguing in the 1970s extended the practice of creating bibliographic records manually by providing a means for such records to be shared, thereby reducing costs and duplication. Even today, libraries achieve bibliographic control by means of sharing manual cataloguing of printed resources through electronic distribution.

For networked resources, however, cataloguing is found most inadequate for describing contents and providing access. To take care of the emerging needs of networked resources, a comprehensive and diverse set of controls is created and upgraded continuously. This set is appropriately called METADATA – the data about data.

13.4.2 Metadata

In 2004, NISO defined MEADATA as 'Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage information objects'. It is machine-understandable and support wide range of operations, and may be categorized as Descriptive metadata, Structural metadata, and Administrative metadata. It is believed that 'if librarians are involved at all, their role with respect to metadata will be vastly different for their old cataloguing role'.

Cataloguing Metadata Tools

The Metadata tools for cataloguing networked resources are too many and many of them are dependent on the nature and type of resources. Typology of Metadata can be mapped in the following way:

- Data Structure: Dublin Core, MODS, CDWA, VRA, LOM
- Data Content: AACR2, RDA, FRBR, CCO, CDPDCMBP
- Data Value: LCSH, AAT, TGN, LCTGM, ULAN, W3CDTF
- Data Format: XML, SGML, MARC
- Data Presentation: ISBD, CSS and/or XSLT

Here we briefly introduce the core sets of Metadata developed particularly for online cataloguing, namely AACR/RDA, and MARC, and two more related standards, namely, ISBD and FRBR.

AACR

Anglo-American Cataloguing Rules is a content standard for bibliographic description and access. It covers not just books, but all other print formats as well. Rules for each category of material specify what fields should be used and what data to include in each field. Text strings were originally intended for printed catalogue cards.

The key principles of AACR are:

- One principle entry per resource
- Catalogue from item in hand
- Chief source of information

AACR Timeline. It all begins with the publication of UK and US editions in 1967. The unified second edition, consistent with ISBD specifications, was published in 1978.

The major breakthrough was achieved in 1997 in Toronto conference on AACR2. Next year the Functional Requirements of Bibliographic Records (FRBR) was brought out. In 2005, Resource Description and Access (RDA) was developed.

AACR 2, in its first part deals with bibliographic descriptions: general rules, resource-specific rules, and rules regarding analytical entries. The second part deals with choice and constructions of access points.

The significance of AACR2 as a metadata tool for cataloguing online resources is fading out due to the following shortcomings:

- *Increasingly complex*
- *Lack of logical structure*
- *Mixing content and carrier data*
- *Hierarchical relationships missing*
- *Anglo-American centric viewpoint*
- *Written before FRBR*
- *Not enough support for collocation*
- *Unclear relationship with MARC Format*

RDA

RDA stands for "Resource Description and Access" the new standard that will replace AACR2.

The Aims of RDA are:

- *Rules should be easy to use and interpret*
- *Be applicable to an online, networked environment*
- *Provide effective bibliographic control for all types of media*
- *Encourage use beyond the library community*
- *Be compatible with other similar standards*
- *Have a logical structure based on internationally agreed principles*
- *Separate content and carrier data*

RDA Timeline

The RDA Prospectus was issued along with draft of some chapters in 2005. During next two years further drafts of chapters on description and access were issued. In 2008 IFLA conference a screenshot demo was presented.

Launch of online product has been scheduled in early 2009.

Use of RDA

The purpose of using RDA is primarily to analyse the knowledge resource being described and determine what is the type of content, what is its carrier form, what other resources it relates, which persons, families or corporate bodies is it related, to what concepts, events and places is it related. RDA rules are for applying to all content types and all media types.

RDA aims to be:

Independent of communication formats for bibliographic and non-bibliographic resources:

- UNIMARC, MARC, MARCXML, MODS/MADS
- DC, EAD, ISBD, VRA, MPEG7

Compatible / better aligned with other similar standards

- Archives: ISAD (G)
- Museums: Cataloguing Cultural Objects

MARC

MARC is a communication and exchange format providing a structure for encoding the content of bibliographic and authority data. The MARC format was developed in the late 1960s as a tagging scheme for exchanging cataloguerecords on magnetic tape. It remains the standard way to represent such data. At present, MARC is steadily being converted (slowly) to modern computing formats, e.g., Unicode, XML.

The followings are the primary reasons for the growing dissatisfaction about MARC format:

- A classic legacy system
- Not designed for computer algorithms
- One record per item (poor links between records)
- Tied to traditional materials and traditional practices
- Not Unicode
- 100 of million records at $100 — $10 billion
- Note that the content is designed to be part of a printed cataloguerecord and is not in a convenient format for computer manipulation.

MARC Timeline

In 1960s, Library of Congress designed MAchine-Readable Cataloguing format for developing database of cataloguerecords for producing printed cards. The British Library developed UKMARC in parallel. In 1970s, variant national formats like AUSMARC, DANMARC, were developed. USMARC brought out 8 material formats, Books, Serial, Maps, etc. In 1977, UNIMARC was developed by IFLA to exchange records between national MARCs. Some important changes have taken place in recent time. Those are:

Expansion of USMARC to a family of formats

Bibliographic, Holdings, Authority, Classification, Community Information

Integration of USMARC bibliographic format and previous 8 formats

Widespread adoption of MARC 21

1997 – USMARC and CANMARC become MARC 21

2003/4 – MARC 21 enhanced by UK proposals; British Library adopts MARC 21

2006/7 – MARC 21 enhanced by German proposals: this will enable libraries to move from MAB to MARC21

Modernizing MARC

- Keep the content of the catalogue record
- Convert to Unicode for representing scripts
- Convert to XML for tagging cataloguing metadata.

ISBD

The International Standard Bibliographic Description (ISBD) is a set of rules produced by IFLA to describe a wide range of library materials within the context of a catalogue. One of the original purposes of the ISBD was to provide a standard form of bibliographic description that could be used to exchange records internationally. This would support IFLA's program of universal bibliographic control.

ISBD Timeline

International Standard Bibliographic Descriptions (ISBD) was developed by IFLA 1969 onwards.

The consolidated edition of the ISBD was published in 2007. It superseded earlier separate ISBDs that were published for monographs, older monographic publications, artographic materials, serials and other continuing resources, electronic resources, non-book materials, and printed music. IFLA's ISBD Review Group is responsible for maintaining the ISBD.

ISBD defined seven areas of description, their sequential order, and associated punctuations.

- Title
- Statement of Responsibility
- Edition
- Resource specific information
- Publication details
- Physical description
- Series information
- Notes and standard identifiers

FRBR

FRBR is a product of the logical progression of thought, initiated one and half a century ago, on how to organise the catalogue in this challenging information age.

FRBR Timeline

Functional Requirements of Bibliographic Records (FRBR) was prepared by IFLA in 1998.

From 1992 to 1995, the IFLA Study Group on Functional Requirements for Bibliographic Records developed an entity relationship model as a generalised view of the bibliographic universe, intended to be independent of any cataloguing code or implementation.

FRBR entity-relationship model defines:

Tasks: find, identify, select, obtain

Resourcerelationships:

work, expression, manifestation, item

- Entities: people, corporate bodies (agents)
- Entities: concepts, objects, events, places

FRBR Aims

The idea behind the FRBR conceptual model is that the catalogueis not seen as a sequence of records, but rather as a network of connected data. FRBR clarifies how catalogues should function, and illuminates what information is of the most value to users of the catalogue.

Self Check Exercise

Note: 1) Write your answer in the space given below.

2) Check your answer with the answers given at the end of this Unit.

4) What is Metadata? What are the metadata tools for cataloguing?

....

....

....

....

13.5 OPAC – ONLINE CATALOGUE INTERFACE

13.5.1 What is OPAC

The acronym, OPAC, stands for *Online Public Access Catalogue*. For all practical purpose, we may define OPAC as computer terminal at user end providing a friendly interface for searching, retrieving and viewing the machine-readable cataloguing data in eye-readable text format. Web 2.0 technology helped decoupling the user interface from the back-end systems that support all internal operations of a library including management of databases and services.

13.5.2 Backdrop

Long back in early 1980s, a librarian predicted that "Survival of Library "as an information agency will be dependent on its ability to redefine its procedures and goals in terms of the bibliographic universe as a whole. In doing so, it will be necessary to place its basic tool, the catalogue, in its proper perspective with other access tools."

Today, the scenario remained as it was, if not worsened. After 25 years, OCLC surveyed 396 college students from six countries, on their use of library resources. The survey revealed that 89 percent of them began their information search on a search engine, and only 1 percent on a library catalogue.

It is, however, realised now that the "persistent problems of the catalogueexist less with its business modules and more with its front-end" as R. David Lankes stated at the ALCTS 50th anniversary conference in June 2007.

13.5.3 We Need a New Interface

In an environment of Web 2.0 technology, the online library catalogues with new interfaces call for immediate attention. An unprecedented amount of research effort is now being put by the library automation vendors, and open-source champions in creating new library interfaces to match with users' expectations.

The Net Gen library users of today are Web-savvy, forward-looking and little intolerant. To attract them toward the library web sites and retain their attention, we must recreate our web interfaces competing the commercial Web. If the interface doesn't respond within seconds, a large percentage of users will click away to other sources.

Library catalogues are lagging behind the commercial and social interfaces we find on the Web, both in terms of looks and functionality. The web users prefer to follow a search model comprising faceted navigation and result clustering. It starts with entering a general term, receiving abundant results, and drilling down to incrementally narrow those results. This contrasts against the structured method still prevailing in library environment, where the search process starts with a pre-formulated Boolean-logic, contrary to direct method of keyword searching. It has been observed in 1997 that the library users, following the general trend in web environment, search online catalogue more often by keyword than not.

The gap between the online library catalogue search experience and search experience in commercial and open-source arena will continue to grow further until we reconstruct OPAC - the user-interface of our online catalogue.

13.5.4 OPAC in Next-Generation Web Catalogue

There are quite a few next-generation library interfaces now available on Internet. Those are still in their early stages of development but already have demonstrated the revolutionised ways libraries can interact with users. These interfaces are easy to use, look impressive, and take sophisticated approach to find information, and put the details on view helpfully. Some of the most innovative projects, like Endeca, Encore, eXtensible Catalogue, Aqua Browser Library, Koha, are now working in a number of forward-looking institutions.

13.5.5 Search Tools

Basic expectations from a web-based interface that it should provide the user with search results in order of relevancy ranking, facets (automatically generated terms that can be clicked to narrow results), corrections of misspellings, and suggestions to alternative search terms. To meet these expectations, the new generation library interfaces offer: relevance ranking, faceted navigation, search result clustering, breadcrumb trails, and a faster, more friendly search environment. Each has its strengths. By critically examining the key concepts, features, and techniques associated with these tools, librarians should be able to identify the essential elements of successful search interface that libraries can adapt to their systems.

13.5.6 Relevance Ranking

Whoever searches Internet often becomes used to the idea of relevance ranking where the best and most interesting items rise up to the top of their hit list. This is because all the giant search engines on web, like Google and Yahoo!, work this way.

Relevance ranking works satisfactorily when comes through careful refinement of the formulas that determine the ordering of results. One must develop skill for achieving good relevance ranking.

With relevance ranking, users may expect to get a large result set. That does not make them upset. They know, by looking up the top of the first page the most relevant information should be found.

In fact, most users do not notice how big the result set is.

In spite of the great advantage of relevance ranking, users may often require to sort the search hits by title, author, date, or by some other elements. Provision of sorting by elements of users' choice is another attractive feature of new OPAC.

13.5.7 Faceted Navigation

A well-constructed faceted navigation scheme will allow users to quickly drill down from a broad set of results to a manageable group of results. The search interface selects the facets from metadata that's associated with the items in the body of information. Names, subject headings, publication dates, and other fields from a structured record provide good opportunities for generating facets.

There has been considerable interest in Faceted Application of Subject Terminology, an alternative way of applying the Library of Congress Subject Headings. This system has been designed to be more amenable to interfaces that employ faceted navigation.

Collections that already have rich metadata, such as library catalogues, lend themselves to faceted navigation. Those with sparse metadata or none at all call for result clustering.

13.5.8 Clustering

Clustering of results, by subject, author, genre format or date, can allow users to easily refine a search, with one click of the mouse. New methods of browsing—via peer recommendations, or through subject taxonomies, or related websites—have emerged.

Result clustering, like faceted navigation, aims to give the user a fast and easy way to narrow results. While faceted navigation usually relies on metadata terms, clustering operates by analysing the raw text of the items in the result set to create labels that represent each group or cluster. From the user's perspective, faceted navigation and clustering look much the same. With both, you click through a succession to progressively narrow the results down to a manageable number.

Using new search technologies, we can meet this expectation. Since it was first popularised by Google, relevance ranking has transformed the way that people search.

13.5.9 Breadcrumbs

Breadcrumb links, also known as contextual links, are a type of navigation aid for Web pages. They provide a textual representation of a site's structure, usually a vertical hierarchy of a site. For example, on e-commerce websites, breadcrumb links often have the following format:

Home > Category > Subcategory > Product

Breadcrumb links can help user take some steps back toward home when needed. A new link appears with each step through the site and persists until the user safely exits. Breadcrumbs are especially helpful in search environments that use faceted navigation. New breadcrumbs show each facet that's been selected to narrow the search.

Since this drill-down approach is often an iterative process, a good presentation of breadcrumbs allows the user to easily back out of an existing facet selection and choose another.

13.5.10 Search Result Data

Key challenges for these new interfaces include the need to extract data from the library service environment and provide up-to-the-minute status information on each item in the library's collection, for example 'on circulation'.

Librarians need to construct search interfaces that include the catalogue of hybrid collections that may include a large portions of the electronic content to which they subscribe. The nature of information requirements and information seeking behaviour both have changed in 21st century, as the figure-1 shows. Some of the new generation OPAC are paving the way toward providing more immediate and seamless access to diverse content collections, giving equal footing to both the digital and print collections. As we are experiencing currently, the medium and form of document is irrelevant to information user. What they need is the information. They do not care much if the source document format is a print or electronic, local or remote, or whether it is a text, video, or audio. Our regenerated OPAC must take care of their need and their information seeking behaviour.

Self Check Exercise

Note: 1) Write your answer in the space given below.

2) Check your answer with the answers given at the end of this Unit.

5) What is Relevance Ranking? How does it help researchers? What else they may need to arrange their search hits helpfully?

....

....

....

....

13.6 ONLINE CATALOGUING UTILITY SERVICES

Libraries invest a substantial proportion of funds to generate and update machine-readable catalogue– a labour-intensive time-consuming process. The job demands, besides cataloguing knowledge and skill, a command on metadata for networked resources.

Online catalogues are generated in two ways: by *original cataloguing* or by *copy cataloguing*. In networking environment, copy cataloguing work includes (1) finding sources of MARC records of target documents, (2) downloading records, (3) eliminating unwanted fields, and (4) adding fields of local relevance. Since cataloguing procedures are carried out programmatically and often supported by utility services, it takes less time, involves less cost and less intellectual effort. On the other hand, original*cataloguing* stands for complete cataloguing of a document from scratch. It is undertaken when downloadable records of corresponding documents are not found.

The objectives of the online cataloguing utilities are to provide ways and means to reduce costs enhance productivity and improve cataloguing workflow.

There are websites offering free catalogue search and download options for records retrieved. LC database is one of them. This gesture helps researchers. However, for library cataloguing, the facility proves disobliging. The utility networks operate at regional and international levels, and extend variety of bibliographic support services and products. Distribution of cataloguing data in MARC format against subscriptions is one common agenda. Top organisations, like Library of Congress and OCLC are among them. Besides the Library of Congress and OCLC, there are hundreds of commercial and no-profit agencies who offer variety of tools and support services for original cataloguing, copy cataloguing, data editing, record validation, MARC conversion, and even for spell checking, link checking, display tools. A selected few are mentioned here group wise with their service profiles.

13.6.1 Search View Print

Book Where

WebClarity Software Inc.'s BookWhere product allows users to simultaneously search and retrieve metadata from over 1900 pre-defined databases located around the globe. Bibliographic and media records can be found in seconds and exported in a wide variety of formats including MARC 21, FINMARC and UNIMARC. www.webclarity.info

FRBR Display Tool - Free

The FRBR Display Tool sorts the bibliographic data found in a set of MARC records into hierarchical displays by grouping the bibliographic data using the "Works," "Expressions," and "Manifestations" FRBR concepts. Possible uses for the FRBR Display Tool include experimenting with the collocation and sorting of search result sets into the FRBR categories to test concepts; and applying FRBR to local data to evaluate its consistency for FRBR-type development. www.loc.gov/marc/frbr/tool.html

MARCView ™

MARCView™ is an easy-to-use program to view, search, and print any MARC 21, USMARC, CanMARC, UNIMARC, or MARCXML bibliographic or authority file. Records are formatted for easy viewing and printing. Navigation to any record in the file is instantaneous. Searches can specify field, subfield, both, or neither. www.systemsplanning.com/marc

MarciveWeb SELECT

Enables librarians to search 10 million records from LC, NLM, NLC, GPO, A/V Access(R), and other sources, and obtain customized MARC 21 bibliographic records, cataloguecards, smart barcode labels, book labels, and MARC 21 authority records. www.marcive.com

13.6.2 Copy Cataloguing

Impact/ONLINE CAT

Impact/ONLINE CAT is a Windows-based cataloguing system which provides users with the ability to search multiple databases of MARC 21 bibliographic, authority and community information records. Features include a MARC Editor with built-in MARC validation, and the ability to download records to the local hard drive. www.autographics.com

Connexion

This is a robust suite of full-service online cataloguing tools and services backed by OCLC's 35+ years of cataloguing experience. Its enhanced features provide unparalleled flexibility for libraries and other allied institutions. Connexion lets you create and edit high-quality bibliographic and authority records, then share them with the entire OCLC cooperative, which benefits libraries around the world.

OCLC-MARC Record Delivery

A variety of WorldCat services generate bibliographic records in OCLC-MARC format. OCLC makes copies of these records available for you to import to your local system to keep your local holdings in synch with your holdings in WorldCat. Downloading records on a regular basis makes it easier for your systems staff to manage and upload them into your OPAC.

Surpass Copycat

Surpass Copycat is a Windows-based Z39.50 copy cataloguing tool that allows users to find and download free MARC records from the Internet. Search multiple libraries simultaneously, such as the Library of Congress, public libraries, medical libraries, state-wide union catalogues and more. Over 100 libraries come pre-configured. Copycat also features "scan and search" that allows the user to simply scan the EAN/ISBN barcode from the back of the book they wish to catalogueto instantly launch a search for that book. www.surpasssoftware.com/copycat.htm

WorldCat Cataloguing Partners

OCLC record delivery in coordination with your vendor orders

Through a collaborative effort with materials vendors, corresponding OCLC MARC records are delivered with the materials you order through participating vendor partners. Additionally, your library's holdings are set automatically in WorldCat.

13.6.3 Converter

MARC RTP - Free

MARC RTP will read files of bibliographic records in MARC format, and convert them to a format that the user designs. The program can also produce a human readable listing or summarize the structure of a file of records. www.loungebythelake.com/marcrtp/

MARConvert™

MARConvert™ handles special problems or unusual requirements in converting records into or out of MARC 21, USMARC, CanMARC, UNIMARC, or MARCXML. It will also convert MARC records to another character set, such as ANSEL, Latin-1, Unicode, or UTF-8. Operates in both interactive mode, and batch mode for converting multiple files. For Windows 95/98/NT/2000/XP. www.systemsplanning.com/marc/mvd.asp

MARCMaker - Free

MARCMaker is developed by the Library of Congress (LC) that generates the MARC record structure from preformatted text. It runs under DOS or Windows 95/98/ME/ 2000. (www.loc.gov/marc/makrbrkr.h tml#download)

(www.loc.gov/marc/makrbrkr.html

USEMARCON Plus - The Universal MARC Record Converter - Free

Like the original UseMARCON program, the USEMARCON Plus program enables libraries to create rules-based systems to convert records between national MARC formats. It also allows users to create and modify rules files, used to achieve MARC conversions, in order to meet specific local requirements. www.bl.uk/services/ bibliographic/usemarcon.html

13.6.4 Validator

MARC Report

MARC Report validates MARC records according to the latest LC and OCLC standards. The validation that is applied is customizable by the user. The program runs either in interactive mode (record-by-record) or in batch mode (producing a report of all problems found). Unique to MARC Report are hundreds of cataloguing cross-checks which check the internal logic of each record, making sure that data elements present in one field do not conflict with those present in another. www.marcofquality.com

Validator ™ Subjects and Names Authority Database

Validator ™ Subjects and Names Authority database includes the complete Library of Congress Subjects and Names Authority Files on CD-ROM. Included are 248,000 subject headings and 4.7 million names records specifying personal and corporate names, series and uniform titles. Validator is versatile and useful to both catalogueers and reference staff. www.att.com/gov/library/

13.6.4 Validator

CILLA

Co-operative of Indic Language Library Authorities

The CILLA service provides quarterly book lists for Bengali, Gujarati, Hindi, Panjabi, Tamil and Urdu materials suitable for public library audiences. This co-operative service enables library authorities to purchase materials more easily, and in confidence, for languages where they do not have a local specialist. The supply of MARC records also enables library authorities to gain efficiencies in cataloguing.

Cataloguing Calculator - Free

The Cataloguing Calculator finds variable and fixed MARC fields (bibliographic and authority data), language codes, geographic area codes, publication country codes, AACR2 abbreviations, LC main entry and geographic Cutter numbers. http:// home.earthlink.net/~banerjek/calculate/

InfoWorks Link Checker

InfoWorks Link Checker is designed for librarians to quickly check URL links in MARC 21 records and works with flat MARC files containing URLs. InfoWorks Link Checker can also check multiple links simultaneously by applying multiple threading technology. www.itcompany.com/linkcheck.htm

InfoWorks Spelling Checker for Database Maintenance

Locates spelling errors and cleans up bibliographic databases. It works with flat MARC files in any integrated library systems. InfoWorks Spelling Checker for Database Maintenance allows users to define which languages and fields to check and provides both batch and interactive checking. It includes a special dictionary for library use and allows users to build custom dictionaries.

www.itcompany.com/checkerd.htm

Self Check Exercise

Note: 1) Write your answer in the space given below.

2) Check your answer with the answers given at the end of this Unit.

6) How machine-readable catalogues are generated in networking environment? How do the utility services help catalogue generation? Name one 'Copy cataloguing' utility service.

....

....

....

13.7 SUMMARY

The theme of this Unit is 'online cataloguing: design and service'. We discussed six different topics touching upon different aspects of online cataloguing. Before going into the details of online cataloguing, we reviewed the development of library catalogue from its early stage up to its latest manifestation in networking environment. This helps students to gain an insight into the objectives of a catalogue, and importance of cataloguing standards. It was then; we take up the functional aspects of online catalogue, as distinguished from electronic catalogue used for library automation, in context of MARC database. Subsequently, the internal structure of a MARC record was examined taking the requirements of bibliographic fields in view. Next, we observed the functionality of few metadata tools in building online catalogues, including AACR2/RDA. Then, OPAC as user interface was discussed in the context of advancing web 2.0 technologies. Last, we observed the requirements of 'copy cataloguing' and 'original cataloguing' and the features of various utilities, offered by network services for generation, conversion, validation of MARC records, and other supports. This gives an exposure to the different tasks associated with cataloguing of networked resources, apart from awareness of the utility services as such.

13.8 ANSWERS TO SELF CHECK EXERCISES

1) The technology of Web 2.0 has opened up alternative ways to develop many different search models for bibliographic database. Quite a few successful catalogue systems already visible online, though still in experimental stage. These products of Web 2.0 technology are commonly referred to as next-generation catalogue. They:

- give the user a simple search interface that allows the user to enter vague, broad, and simple searches.
- allow the user to drill down through the large result list, narrowing it down by whatever criteria they choose, until it is as precise as they want.

- sort the results list so that the most relevant items are at the top of the list
- tolerate misspellings and unusual word choices in the user’s search.

The list shows what users desire to get, and what the next-generation catalogues can meet. Next-generation catalogues give users the same tools they already enjoy on websites. Since library databases are generally built in compliance with MARC, they look beyond ILS and will have little problem in working with a variety of systems.

2) MARC is an acronym stands for MAchine Readable Catalogue. MARC standards consist of the MARC formats, which are standards for the representation and communication of bibliographic and related information in machine-readable form. The key to the machine-readability is this common record format. MARC follows a physical structure and a set of control mechanism for enabling the machine to identify and process the data elements, whenever needed. This physical data structuring scheme, originally developed as a carrier of MARC data, is now being used as a standard, nationally (ANSI Z39.2) and internationally (ISO 2709), by other communication formats for bibliographic information interchange, like UNIMARC, CCF, national MARCs, etc. They all are implementations of ANSI Z39.2 / ISO 2709 standard.

The LC MARC found its way to US MARC, as an acknowledged national format for bibliographic communications, and after that evolved into MARC 21. MARC 21 is not a new format but a harmonized edition of USMARC, CAN/MARC, UKMARC and AUSMARC brought out as a consolidated scheme. Although, one finds there little change content wise, MARC 21 differs significantly in its vision. It tends to take issues beyond national preferences, and to negotiate with the up-coming events as well. Being its focus shifted from geographic to temporal zone, MARC 21 appears as a MARC version for the 21st century.

3) **Field Tag**

MARC reserves a number of 3-digit codes, or Field Tags, each represents a particular type of data. The 3-digit codes are, in fact, abbreviated form of the field names used for describing bibliographic units. Field Tag provides bibliographic format with flexibility. This is an indispensable feature for recording bibliographic elements since many data fields, e.g. Title, Author, etc. *vary in length*. Field Tag, being potentially a repeatable device, supports *multiple occurrences* of any field specified as ‘Repeatable’ in MARC 21 documentation. When a field, e.g. Tag 700, repeats twice it will produce two Alternate Authors fields. The nature of the data content of a field type determines repeatability (R) /non-repeatability (NR) of fields. For example, a bibliographic record supports only one non-repeatable (NR) main entry field 100 for Personal Author.

100 1_ $a Spilsbury, Richard, $d 1963-	NR	
020 __ $a 0516242946 (lib. bdg.)	R	
020 __ $a 0516278819 (pbk.)		
650 _1 $a Tigers.	R	
650 _1 $a Endangered species.		

Examples of Repeatable Field Tags

4) **Metadata**

In 2004, NISO defined MEADATA as ‘Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage” information objects’. It is machine-understandable and support wide range of operations, and may be categorised as Descriptive metadata, Structural metadata, and Administrative metadata. It is believed that ‘if librarians are involved at all, their role with respect to metadata will be vastly different from their old cataloguing role’.

Cataloguing Metadata Tools

The Metadata tools for cataloguing networked resources are too many and many of them are dependent on the nature and type of resources. Typology of Metadata can be mapped in the following way:

- Data Structure: Dublin Core, MODS, CDWA, VRA, LOM
- Data Content: AACR2, RDA, FRBR, CCO, CDPDCMBP
- Data Value: LCSH, AAT, TGN, LCTGM, ULAN, W3CDTF
- Data Format: XML, SGML, MARC
- Data Presentation: ISBD, CSS and/or XSLT

Here we briefly introduce the core sets of Metadata developed particularly for online cataloguing, namely AACR/RDA, and MARC, and two more related standards, namely, ISBD and FRBR.

5) **Relevance Ranking**

Whoever searches Internet often becomes used to the idea of relevance ranking where the best and most interesting items rise up to the top of their hit list. This is because all the giant search engines on web, like Google and Yahoo!, work this way.

Relevance ranking works satisfactorily when comes through careful refinement of the formulas that determine the ordering of results. One must develop skill for achieving good relevance ranking. With relevance ranking, users may expect to get a large result set. That does not make them upset. They know, by looking up the top of the first page the most relevant information should be found. In fact, most users do not notice how big the result set is.

In spite of the great advantage of relevance ranking, users may often require to sort the search hits by title, author, date, or by some other elements. Provision of sorting by elements of users’ choice is another attractive feature of new OPAC.

6) Online catalogues are generated in two ways: by *original cataloguing* or by *copy cataloguing*. In networking environment, copy cataloguing work includes (1) finding sources of MARC records of target documents, (2) downloading records, (3) eliminating unwanted fields, and (4) adding fields of local relevance. Since cataloguing procedures are carried out programmatically and often supported by utility services, it takes less time, involves less cost and less intellectual effort. On the other hand, original *cataloguing* stands for complete cataloguing of a document from scratch. It is undertaken when downloadable records of corresponding documents are not found.

The objectives of the online cataloguing utilities are to provide ways and means to reduce costs enhance productivity and improve cataloguing workflow.

Connexion Utility Service

This is a robust suite of full-service online cataloguing tools and services backed by OCLC's 35+ years of cataloguing experience. Its enhanced features provide unparalleled flexibility for libraries and other allied institutions. Connexion lets you create and edit high-quality bibliographic and authority records, then share them with the entire OCLC cooperative, which benefits libraries around the world.

13.9 KEYWORDS

AACR2 : Anglo-American Cataloguing Rules [AACR2 - 2nd Edition (1978), AACR2R - Revised 2nd edition (1988)]

ALA : American Library Association

BL : British Library

DUBLIN CORE : Dublin Core is a 15-element metadata element set intended to facilitate discovery of electronic resources

FRBR : Functional Requirements for Bibliographic Records—or FRBR is a conceptual entity-relationship model developed by IFLA that relates user tasks of retrieval and access in online library catalogues and bibliographic databases from a user's perspective.

IFLA : International Federation of Library Associations and Institutions

ISBD : International Standard Bibliographic Description

LA : Library Association

LIBRARY 2.0 : Application of interactive, collaborative, and multi-media web-based technologies to web-based library services and collections

OCLC : Online Computer Library Centre Inc.

OPAC : Online Public Access Catalogue

MARC : An acronym stands for Machine Readable Catalogue. MARC standards consist of the MARC formats, which are standards for the representation and communication of bibliographic and related information in machine-readable form

MARC21 : The "harmonization" of USMARC and CAN/MARC; maintained by the Network Development and MARC Standards Office of the Library of Congress.

METADATA : Metadata is information about an informational data about data resource, be that a document (such as a webpage), image, dataset or other resource.

RDA : RDA stands for "Resource Description and Access" the new standard that will replace AACR2

UNICODE : Unicode is a computing industry standard allowing computers to consistently represent and manipulate text expressed in most of the world's writing systems.

UNIMARC : Universal MARC Format

WEB 2.0 : Refers to a perceived second generation of web development and design, that facilitates communication, secure information sharing, interoperability, and collaboration on the World Wide Web.

13.10 REFERENCES AND FURTHER READING

Kehal, Harbhajan S and Varinder P. Singh. *Digital Economy: Impacts, Influences, and Challenges.* Hershey, PA : Idea Group Publishing, 2005. Print.

Drake, Miriam A. "Language, Arts and Disciplines".*Encyclopedia of Library and Information Science*: Pub-Zoo. New York: Marcel Dekker, 2003. Print.

Kesselman, Martin Alan and Irwin Weintraub. *Global Librarianship.* New York: Marcel Dekker, 2004 . Print.

Mukhopadhyay, Asok. *Guide to MARC 21for Cataloguing of Books and Serials, with Functional Definitions, Examples, Working Resources.* New Delhi: Viva, 2007. Print.

Verheul, Ingeborg. *Networking for Digital Preservation: Current Practice in 15 National Libraries* . Munchen : IFLA/Saur: 2006. Print.

Joachim, Martin D. *Languages of the World: Cataloguing Issues and Problems.* Chicago: University of Chicago Press, 1994. Print.

Casey, Michael E. and Laura C. Medford. *Library 2.0: A Guide to Participatory Library Service.* N.J.: Information Today, 2007. Print.

Grosch, Audrey N. *Library Information Technology and Networks.* New York: Marcell Dekker, 1995. Print.

Fritz, Deborah A. and Richard J. Fritz. *MARC 21 for Everyone: A Practical Guide.* Chicago: American Library Association, 2003. Print.

UNIT14 OVERVIEW OF WEB INDEXING, METADATA, INTEROPERABILITY AND ONTOLOGIES

Structure

14.0 OBJECTIVES

After reading this Unit, you will be able to:

- define the meaning and need of web indexing;
- explain the role, usage and importance of metadata;
- define is ontology and its importance in web parlance;
- explain interoperability and various methods of interoperability; and
- discuss protocols for interoperability.

14.1 INTRODUCTION

Index is a tool that has been in use for a long time to locate information. It is a list of key words or terms that supplement a document at the end of text for fruitful navigation and browsing. An index not only provides a chance to highlight content and provide a bird's-eye-view to the document, it also helps to identify the inconsistencies and improve upon content of the document for the author. The Web has emerged as an enormous source of information with a lot of chaotic information content also. Structurally, it is a collection of websites hosted at different domains round the globe. Websites can be defined as sources of information, consisting of individual webpages. Index to this content is available at individual level (through websites) and at global level (through search engines).

14.2 WEB INDEXING

Web index is a tool used for searching web documents like, individual websites or collections of web sites or collections of webpages and so on. It is a browsable list of terms or sections leading towards further reading/resources to the desired topic or subject. Sitemap is an example of a web index.

Indexing is an intellectual activity where the indexer determines what concepts are worth indexing. The entries and arrangement of these entries are equally important. There is a view that web indexing as well as traditional indexing is best done by individuals, skilled in the art of indexing. It requires imagination and formal knowledge of the subject.

However, there are automated ways of doing web indexing. Search engines use a program called spider or crawler to extract the search terms from the individual webpages. These spiders or crawlers collect the terms and store the terms inside a local database of a search engine and use it as a search index. These terms are either extracted from the 'meta tag' or from the contents of the webpage.

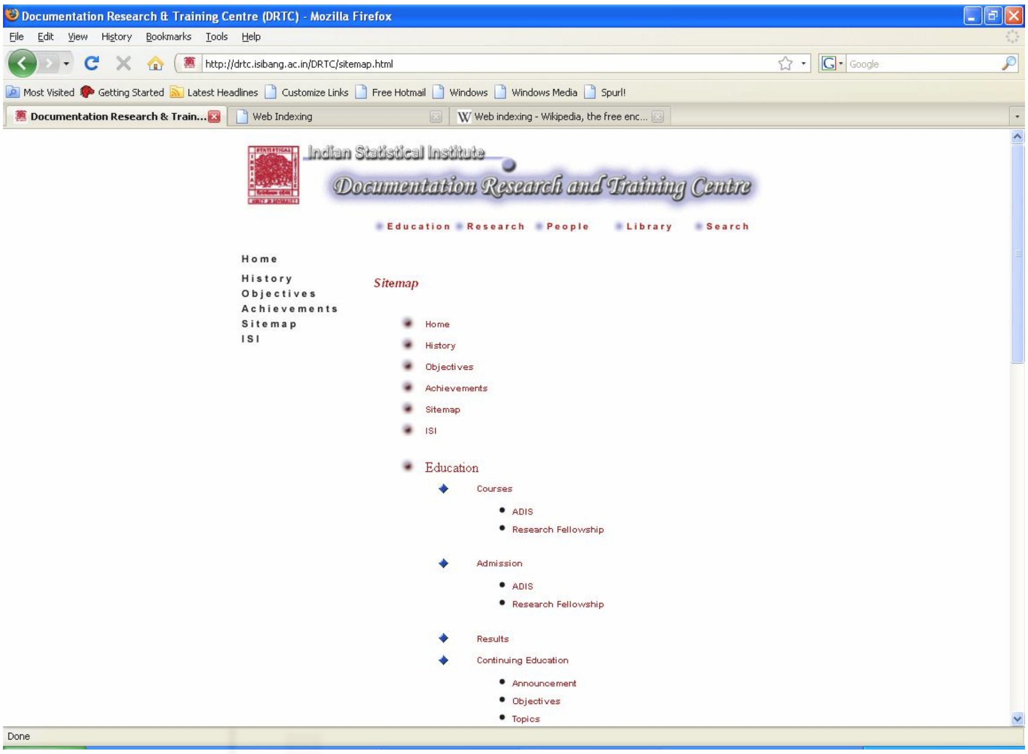
A web index is often a browsable list of entries from which the user makes selections. The index may not be displayed to the user but the user may retrieve information by just typing her/his query into a search box. A website A-Z index is a kind of web index that resembles an alphabetical back-of-the-book style index, where the index entries are hyperlinked directly to the appropriate web page or page section, rather than using page numbers.

According to British Indexing Standard (BS3700:1988) "Web index is a systematic arrangement of entries designed to enable users to locate information in a document."

Web indexing is the process of creating index manually or mechanically for the content of web documents. It includes back-of-book-style indexes for individual websites or an Intranet. It may further include the creation of metadata based keywords to provide a more useful vocabulary for the Internet or onsite search engines. With the increase in the number of e-journals, web indexing has also become important for publishing houses.

14.2.1 Concept

Index is a tool to help users locate information quickly and easily. Often it is understood as list of terms or phrases. But it is something beyond that. It brings like concepts together by grouping and creates a concept map in the mind of user about the document. Similarly, web index is a tool to locate easily and quickly the information on a website. A site map of a website is an index. Normally, indexes are used for web browsing. The terms in the sitemap of a website are directly hyperlinked to the topical web page or to the topic itself within the webpage. It performs the following three important tasks:


Documentation Research & Training Centre (DRTC) - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://drtc.isibang.ac.in/DRTC/sitemap.html
Google
Most Visited Getting Started Latest Headlines Customize Links Free Hotmail Windows Windows Media Spurl!
Documentation Research & Train...
Web Indexing
Web indexing - Wikipedia, the free enc...
Done

Indian Statistical Institute

Documentation Research and Training Centre

Education Research People Library Search

Home
History
Objectives
Achievements
Sitemap
ISI

Sitemap

- Home
- History
- Objectives
- Achievements
- Sitemap
- ISI
- Education
  - Courses
    - ADIS
    - Research Fellowship
  - Admission
    - ADIS
    - Research Fellowship
  - Results
  - Continuing Education
    - Announcement
    - Objectives
    - Topics

Fig. 14.1: Example of Sitemap Index of DRTC website

Source: http://drtc.isibang.ac.in/DRTC/

1) It describes the relationships among subjects. These include hierarchical and other relationships that exist amongst the subjects.

2) Index once stored in the database can be used as a source of meaningful metadata for searching. The search engines use index terms stored in such a type of database. An index provides visibility to all the available literature. Sometimes a user makes a search but cannot express the exact search phrase. In such a case, a good index brings material related to what s/he is looking for, with the help of related terms or concepts.

3) It is not only useful from the searchers' point of view but also for the point of view of the author. An index focuses on the content of the document and demonstrates the inconsistencies about the treatment of the topics. Hence, it is an aid for authors to review the writing for completeness of content.

Web index speeds up the browsing by presenting a browsable conceptual map about the content of the document. It also facilitates searching if used with proper search algorithm and search tools.

Self Check Exercises

Note: 1) Write your answer in the space given below.

2) Check your answer with the answers given at the end of this Unit.

1) Discuss the need of index in web parlance.

....

....

....

....

14.2.2 Types of Web Indexes

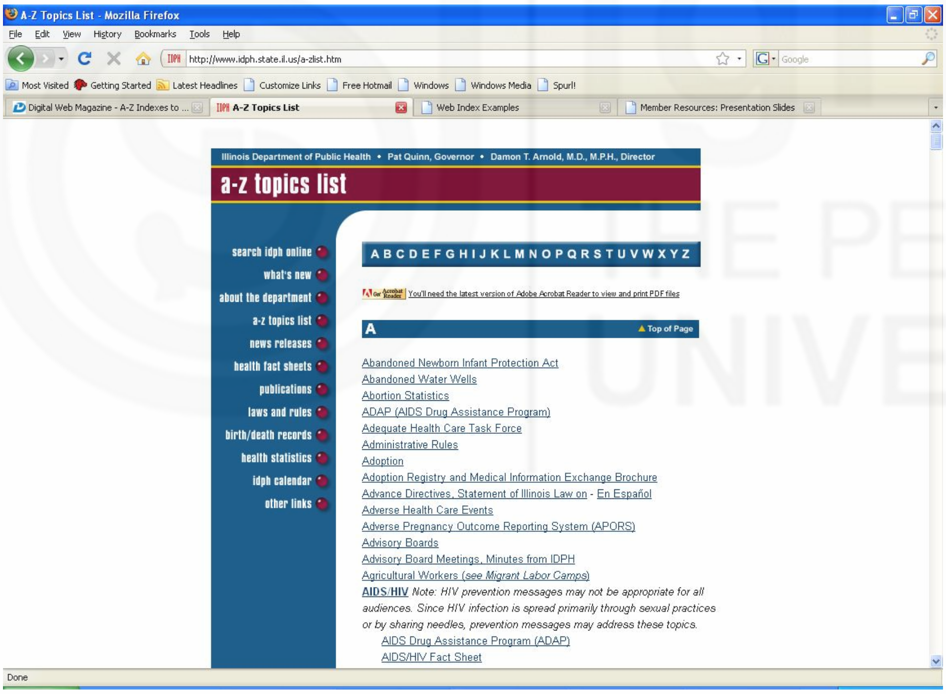
Web indexes are of following types:

Hyperlinked A-Z indexes

Hyperlinked index is a kind of back-of-book-index. It is arranged alphabetically A-Z. Normally, in a back-of-book index the terms or phrases are listed with the appropriate page number or section number. In web environment, A-Z index is a Webpage or a group of pages. Each entry in the Webpage is hyperlinked to a topic or to be more precise to the anchor tag of the resource. The list may also contain synonymous terms linked to the same resource.

If, hyperlinked A-Z Index is prepared manually the rendering of search term rectifies several searching problems like, spelling mistakes, spelling variants, singular plural and so on. There are following visible advantages with the hyper-linked A-Z Index.

- A-Z indexes are most user-friendly.
- The browsable nature of the index can reveal other topics of interest to the user.
- Index entries can link to precise points within a Webpage through the use of named anchor links.
- An A-Z index can enhance the search engine optimisation ranking of the website.


A-Z Topics List - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.idph.state.il.us/a-zlist.htm
Google
Most Visited Getting Started Latest Headlines Customize Links Free Hotmail Windows Windows Media Spurl!
Digital Web Magazine - A-Z Indexes to ...
A-Z Topics List
Web Index Examples
Member Resources: Presentation Slides
Done

Illinois Department of Public Health • Pat Quinn, Governor • Damon T. Arnold, M.D., M.P.H., Director

a-z topics list

search idph online
what's new
about the department
a-z topics list
news releases
health fact sheets
publications
laws and rules
birth/death records
health statistics
idph calendar
other links

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Get Acrobat Reader You'll need the latest version of Adobe Acrobat Reader to view and print PDF files

A ▲ Top of Page

Abandoned Newborn Infant Protection Act
Abandoned Water Wells
Abortion Statistics
ADAP (AIDS Drug Assistance Program)
Adequate Health Care Task Force
Administrative Rules
Adoption
Adoption Registry and Medical Information Exchange Brochure
Advance Directives, Statement of Illinois Law on - En Español
Adverse Health Care Events
Adverse Pregnancy Outcome Reporting System (APORS)
Advisory Boards
Advisory Board Meetings, Minutes from IDPH
Agricultural Workers (*see Migrant Labor Camps*)
AIDS/HIV *Note: HIV prevention messages may not be appropriate for all audiences. Since HIV infection is spread primarily through sexual practices or by sharing needles, prevention messages may address these topics.*
AIDS Drug Assistance Program (ADAP)
AIDS/HIV Fact Sheet

Fig. 14.2: Hyperlinked A-Z Index

Source: http://www.idph.state.il.us/a-zlist.htm

Meta-tag Keyword Indexing

Meta-tag is used in HTML (Hypertext Markup Language) documents for page description, keywords and other metadata. It is used in the header section of the web. The content of the tag is not visible on web browser. Metadata is normally referred as ‘data about some object’. The object could be anything. The data about the object reflects the properties of the object. Some examples of objects in a bibliographic database are Title, Author, Place, Publisher etc.. Rendering of meta-tag in an HTML document is done as follows:

```
<html>

      <head>

            <title>Title of the webpage</title>

            <meta name="title" content="Indira Gandhi National
      Open University" />

            <meta name="author" content="Aditya Tripathi" />

      </head>
```

Fig. 14.3: Metatag in HTML Document

Note: Apart from describing the webpage, meta-tag is used for number of other purposes like, redirection from one page to other, handling the robot of search engines etc.

Robot is a program used by search engine in order to extract data from the web pages so that pages can be searched using the search engine's search interface. Robot is also known as Crawler or Spider. Following are the names of robots used by popular search engines,

Table 14.1: Robots used by Search Engines

Search engine	Robot
Google	Googlebot
Yahoo	Slurp
MSN	MSNbot

If the search engine is compliant with a metadata schema for example, Dublin Core then robot of the search engine extracts the metadata easily given on webpage and stores in the database of search engine.

Each metadata entry of webpage is used as an index term or phrase, further broken into keywords. With such an index context of the keywords or phrase is also extracted with the name of meta-tag. This kind of index is known as meta-tag keyword indexing.

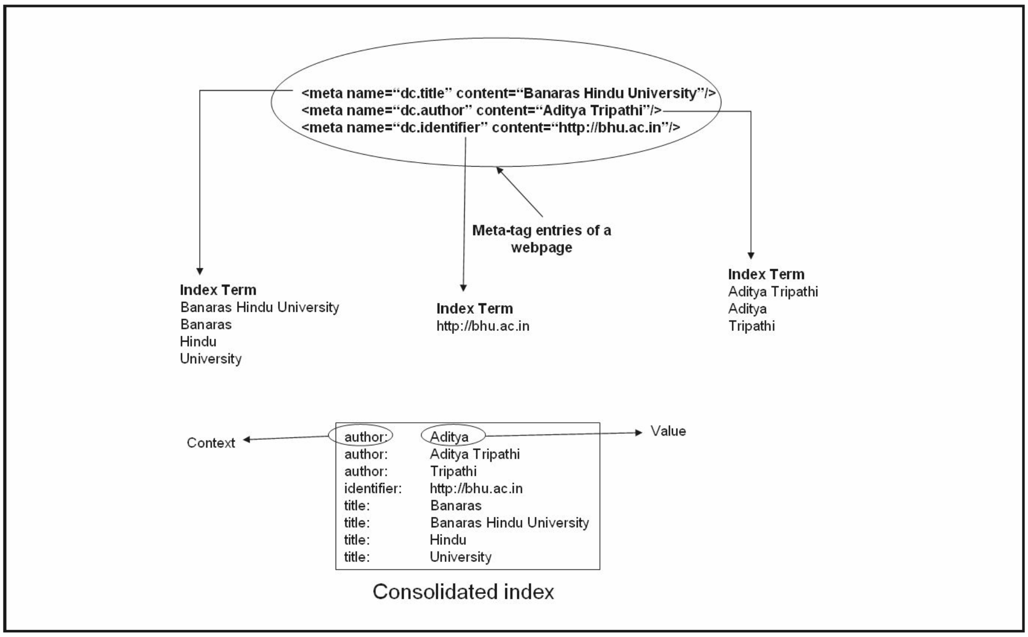

<meta name="dc.title" content="Banaras Hindu University"/>
<meta name="dc.author" content="Aditya Tripathi"/>
<meta name="dc.identifier" content="http://bhu.ac.in"/>
Meta-tag entries of a webpage
Index Term
Banaras Hindu University
Banaras
Hindu
University
Index Term
http://bhu.ac.in
Index Term
Aditya Tripathi
Aditya
Tripathi
Context
Value
author: Aditya
author: Aditya Tripathi
author: Tripathi
identifier: http://bhu.ac.in
title: Banaras
title: Banaras Hindu University
title: Hindu
title: University
Consolidated index

Fig. 14.4: Meta-tag Keywords Indexing

The advantages of metadata index is as follows:

- Context is preserved with the search term, which leads to precision in the search results. For example, documents on Ranganathan and documents by Ranganathan can be easily differentiated.
- Terms are extracted automatically from the web pages through a robot. This kind of index is useful in automatic indexing.

Keyword Creation for Search Engine Optimisation

Search engines look for search term which is queried in its database and fetches the result. If a document appears first in the order of search result then it is said that the page has better visibility. The ordering of results from the search engine is known as ranking. Webmaster who designs and maintains the website attempts to have best visibility in the search result. In order to achieve the visibility it is required to render relevant terms in the meta-tag element of webpages. This is known as Search Engine Optimization (SEO). Search Engine Optimization is of two types,

- White Hat SEO
- Black Hat SEO

White Hat SEO

Often search engines provide guidelines for the webmasters or content developers to have better visibility and ranking of website. For example, Google and Yahoo both provide guidelines for webmasters or content developers. If a webmaster follows these guidelines her/his website will get better visibility. In other words, White Hat SEO is a kind of web development technique that promotes accessibility.

Black Hat SEO

In order to improve upon ranking of webpage in search result many webmasters resort to unfair means of using heavy number of keyword count within the page. This is known as ‘Spamdexing’. They often put more number of keywords in page with same colour as the background. Because of this keywords are not visible for human eye where as robots can read them. Similarly, webmasters play trick and present different webpage for search engine and human accesses to the website, deceiving search engines. This is known as cloaking. Search engines attempt to find out such kind of unethical methods of improving ranking and often lead to banning these websites.

Taxonomies/Categories

Taxonomy refers to the abstract structure of a subject. It is also referred as subject-based classification. Taxonomy typically displays the hierarchical structure of various components or sub-disciplines.

In a taxonomy like terms or subjects are grouped together so that finding the correct term becomes easy. It is used to identify the subject of the document. A typical, taxonomy is given as follows:

Library Science

+Classification

-Enumerated classification

-Analytico-synthetic classification

+Cataloguing

-Descriptive Cataloguing

-Simplified Cataloguing

```
Library Science
     +Classification
          -Enumerated classification
          -Analytico-synthetic classification

     +Cataloguing
          -Descriptive Cataloguing
          -Simplified Cataloguing
```

Fig. 14.5: Taxonomy

Use of taxonomy for the purpose of indexing, facilitates grouping the like objects or documents together. It displays all the objects or documents which belong to one category. Taxonomy is a kind of a controlled vocabulary. Hence, it can be also used as authority control.

Thesauri

Thesauri are also a kind of controlled vocabulary. Thesaurus is taxonomy with enhanced functionalities. Thesaurus demonstrates the relation of terms with respect to Broader Terms (BTs), Narrower Term (NTs), Related Terms (RTs), Synonymous Terms (SNs), Usage, Top Term (TT) and so on. The terms in a thesaurus are usually listed alphabetically.

Table 14.2: Terms in a Thesaurus

Broader Term	Broader in scope than the terms that are subordinate to it in a thesaurus hierarchy
Narrower Term	More specific concept than its parent term in the thesaurus hierarchy
Related Terms	A Preferred Term linked to another preferred term conceptually but not hierarchically
Top Term	The most general terms in a thesaurus hierarchy
Synonymous Terms	Term carries same meaning

Following example is taken from the thesaurus on agriculture, AGROVOC.

```
Pollution
     NT: Acid deposition
     NT: Air pollution
     NT: Nonpoint pollution
     NT: Sediment pollution
     NT: Water pollution
     RT: Environmental degradation
     RT: Pollutants
     RT: Pesticides
```

Fig. 14.6: Terms used in AGROVOC

Source: AGROVOC

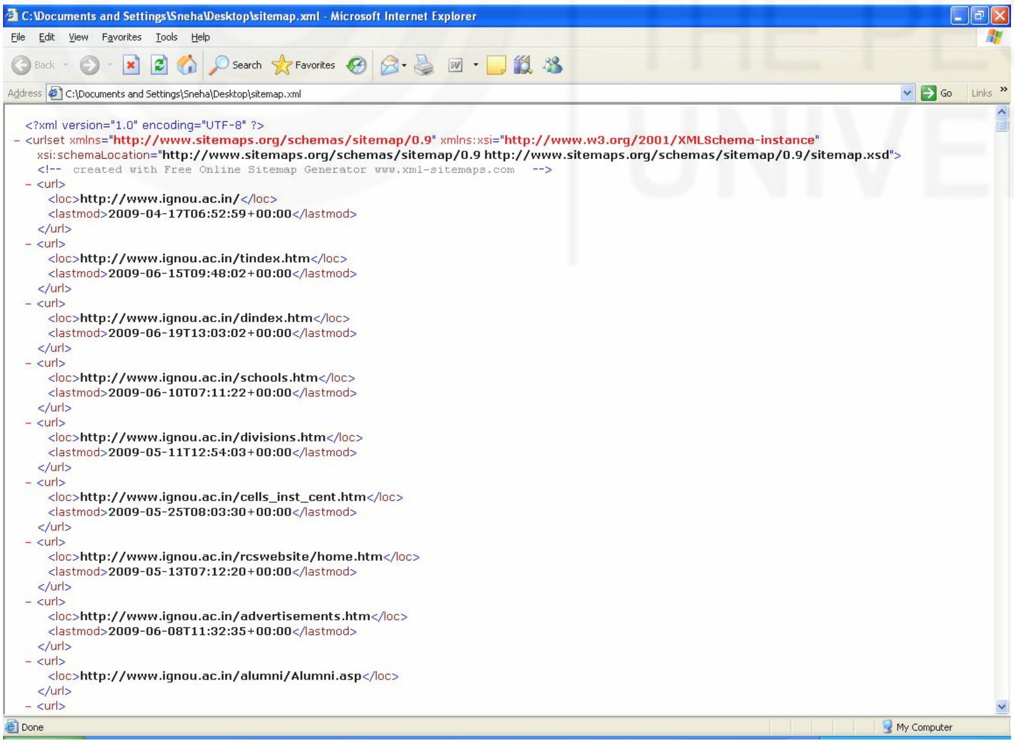
Using the thesaurus has the following benefits:

- An index with the help of thesaurus brings standardisation in rendering the terms.
- Use of thesaurus brings lot of relations like BT, NT, RT and so on which further leads in search refinement.
- If thesaurus is bilingual or multilingual, it can be used for text translation or cross-lingual information retrieval.
- Thesaurus can be used as authority control.

Sitemaps

A good website should be supplemented by good sitemap then only it is said to be complete. A sitemap displays structure of website and the flow of information in it. Hence, sitemap is a document detailing the various pages on a website and their links to each other. This helps the visitors both in finding and searching the pages. The use of the sitemap is to enhance browsing. Though this is the original idea of preparing site map but in due course of time the use of sitemap has changed a lot. Now it is used for exposing the hidden and dynamic content to the search engines using a 'sitemap index' file.

Sitemap index is an XML file (Extensible Markup File), which is prepared in a particular format and submitted to a search engine. There are programs available over the Internet which generates XML based sitemap index. This file can be downloaded and kept in the root directory, when search engine's crawler visits the website it picks up *sitemap.xml* file. Otherwise, it can be submitted directly to the search engines like Yahoo or Google. There may be some difference among the format of *sitemap.xml* file depending upon the search engines.


C:\Documents and Settings\Sneha\Desktop\sitemap.xml - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address C:\Documents and Settings\Sneha\Desktop\sitemap.xml
Go
Links
Done
My Computer

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.sitemaps.org/schemas/sitemap/0.9 http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd">
    <!-- created with Free Online Sitemap Generator www.xml-sitemaps.com -->
  - <url>
      <loc>http://www.ignou.ac.in/</loc>
      <lastmod>2009-04-17T06:52:59+00:00</lastmod>
    </url>
  - <url>
      <loc>http://www.ignou.ac.in/tindex.htm</loc>
      <lastmod>2009-06-15T09:48:02+00:00</lastmod>
    </url>
  - <url>
      <loc>http://www.ignou.ac.in/dindex.htm</loc>
      <lastmod>2009-06-19T13:03:02+00:00</lastmod>
    </url>
  - <url>
      <loc>http://www.ignou.ac.in/schools.htm</loc>
      <lastmod>2009-06-10T07:11:22+00:00</lastmod>
    </url>
  - <url>
      <loc>http://www.ignou.ac.in/divisions.htm</loc>
      <lastmod>2009-05-11T12:54:03+00:00</lastmod>
    </url>
  - <url>
      <loc>http://www.ignou.ac.in/cells_inst_cent.htm</loc>
      <lastmod>2009-05-25T08:03:30+00:00</lastmod>
    </url>
  - <url>
      <loc>http://www.ignou.ac.in/rcswebsite/home.htm</loc>
      <lastmod>2009-05-13T07:12:20+00:00</lastmod>
    </url>
  - <url>
      <loc>http://www.ignou.ac.in/advertisements.htm</loc>
      <lastmod>2009-06-08T11:32:35+00:00</lastmod>
    </url>
  - <url>
      <loc>http://www.ignou.ac.in/alumni/Alumni.asp</loc>
    </url>
  - <url>
```

Fig. 14.6: Sitemap.xml of IGNOU website

Sitemaps are important and beneficial at places where:

- some part of website is not visible due to use of dynamic scripts like Java pages or PHP pages, or

- pages where rich Ajax or Flash content is used.

Self Check Exercises

Note: 1) Write your answer in the space given below.

2) Check your answer with the answers given at the end of this Unit.

2) Discuss the different types of web indexes.

....

....

....

....

14.3 METADATA

14.3.1 Concept

Metadata is information about an informational resource, the document can be a webpage, image, dataset or other resource. Metadata is valuable towards storage and retrieval of documentary resources. Structured metadata make objects easily discoverable. In library parlance catalogue is known as metadata which provides descriptive information about an object or resource available in library. The resource can be physical or electronic. Metadata is a tool used to locate the object or document. There are various metadata schemas some are as follows:

- Anglo American Cataloguing Rule 2 (AACR2)
- MARC21
- Government Information Locator Service (GILS)
- Encoded Archive Description (EAD)
- Dublin Core

Hence, metadata is "data about data". It is structured set of data which describes the various characteristics of an object or document.

The most commonly used metadata schema is Dublin Core Metadata Initiative (DCMI) over the Internet. The standard is developed and maintained by DCMI and DCMI Task Groups. There are 15 elements given in Dublin Core. Apart from these 15 elements there are other metadata set vocabularies which should be used with 15 elements. The following example shows Dublin Core Metadata Record.

Table 14.3: Dublin Core Metadata Record

Creator	Aditya Tripathi
Publisher	Documentation Research & Training Centre
Identifier	http://drtc.isibang.ac.in
Subject	Library and Information Science
Format	txt/html
Language	English
Rights	Indian Statistical Institute

Purpose

Metadata is used for various purposes. These are to:

- retrieve a document;
- define the structure of document and its future maintenance;
- store the preservation conditions; and
- preserve the additional information regarding handling and usage of a document.

14.3.2 Types

As discussed above, it is common practice to use metadata for easy retrieval. But application of metadata has much more role to play in an electronic environment. Based on their roles, metadata are classified in the following types:

- Administrative metadata;
- Technical metadata;
- Structural metadata;
- Descriptive metadata; and
- Preservation metadata.

Administrative Metadata

When a document is created there are several kinds of information also generated with it. These information are valid and useful during the whole life span of the document. These data are stored in as administrative metadata. Administrative metadata is related with the life cycle of the document. It includes information regarding serials in the digital environment concerning:

- Ordering
- acquisition
- maintenance
- licensing
- rights
- ownership and
- provenance

Out of the above information 'rights' and 'digital provenance' are very important.

Technical Metadata

The technical metadata stores information regarding the file type and associated content type and how it should be rendered. It stores information regarding how the bytes should be read or in other words how the file should be read. Apart from this it also stores information regarding size or the extent of the file.

This information is very useful for playing the file. Further it is also useful for digital preservation particularly for migration and refreshing of the document. Technical metadata is helpful in checking the intactness of the object.

Structural Metadata

Structural metadata or structural map of an object explains different components and their role. This handles various sections and sub-sections of the documents and their corresponding relations and roles.

For example, structure of a book is defined as follows:

```
<mets:structMap TYPE="physical">

  <mets:div TYPE="book" LABEL="Martial Epigrams II">

      <mets:div TYPE="page" LABEL="Blank page">

      </mets:div>

      <mets:div TYPE="page" LABEL="Page i: Half title page">

      </mets:div>

      <mets:div TYPE="page" LABEL="Page ii: Blank page">

      </mets:div>

      <mets:div TYPE="page" LABEL="Page iii: Title page">

      </mets:div>

      <mets:div TYPE="page" LABEL="Page iv: Publication info">

      </mets:div>

      <mets:div TYPE="page" LABEL="Page v: Table of contents">

      </mets:div>
```

Fig. 14.7: Example of Structural Metadata

Descriptive Metadata

The metadata used for describing the documents in library is descriptive metadata. AACR2 or MARC21 are good example of descriptive metadata. In library parlance we call it descriptive cataloguing. Descriptive metadata stores information regarding title, author, place, and publisher and so on. This metadata set is important for identifying and locating the documents. For document location over the web, Uniform Resource Identifier (URI) is used. In case of traditional documents in library it is call number of the document where as for web document MARC21 defines field 856 for document location. Dublin core metadata elements have a field called *Identifier* used for document location.

Preservation Metadata

One of the most important metadata set used for digital longevity is preservation metadata. Digital preservation is process of increasing the longevity of documents from physical deterioration. The deterioration of digital objects is against time, technology, media and transfer. In order to secure the document and its original features, libraries, archives and museums need some kind of documentation in the form of metadata. Preservation metadata stores the preservation conditions of a document and its original features at the time of its digital provenance.

PREMIS (Preservation Metadata: Implementation Strategies) is the standard metadata set used for digital preservation purpose. It is joint venture of OCLC and RLG (Research Library Group). The working group comprised of experts of international repute working in digital preservation and metadata usage. The working group developed a core set of implementable preservation metadata and implementation guidelines in terms of creation, management and use of metadata. The working group came out with a set of Data Dictionary for Preservation Metadata. Current version of PREMIS is 2.0. However, PREMIS does not concentrate on descriptive metadata set because it is domain specific and second there are many descriptive metadata schemas available for use.

The preservation metadata is useful for following purposes:

- Supporting the viability, renderability, understandability, authenticity, and identity of digital objects in a preservation context;
- Representing the information most preservation repositories need to know to preserve digital materials over long-term;
- Emphasising "implementable metadata": rigorously defined, supported by guidelines for creation, management, and use, and oriented toward automated workflows; and
- Embodying technical neutrality: no assumptions made about preservation technologies, strategies, metadata storage and management, etc.

Self Check Exercises

Note: 1) Write your answer in the space given below.

2) Check your answer with the answers given at the end of this Unit.

2) Discuss the different types of metadata.

....

....

....

....

14.4 ONTOLOGY

14.4.1 Concept

Ontology is derived from the Greek *Onto* (being) and *Logia* (written or spoken discourse). It is part of metaphysics, branch of philosophy. Ontology studies existence of entities and their relationships. The relationship is derived due to grouping the entities based on formed groups. These groups are formed due to likeness or similarities of characteristics or attributes of individual entities. The relationship is depicted in the form of hierarchy and subdivisions. In other words, it is conceptualisation of world, based on entities and their mutual existence as it is studied in Philosophy.

However, ontology is also studied in computer science and information science. In computer science, ontology is the formal representation of a concept or a set of concepts within a specific domain of knowledge and their relationships. For example, Organisms are classified in two categories i.e. Plantae and Animalia. Then Animalia is further classified into Chordata and Non-Chordata. Chordata is further classified into Protozoa, Coelenterate and so on. This is taxonomy of animal kingdom.

One of the important parts of ontology is definition of classes and their properties. Class represents a group of concepts or objects having same kind of properties. Properties can be defined as distinguishing features for identification of a class or an individual object. Hence, it can be stated that the use of ontology is the use of classification for web documents. The term classification is also known as taxonomy.

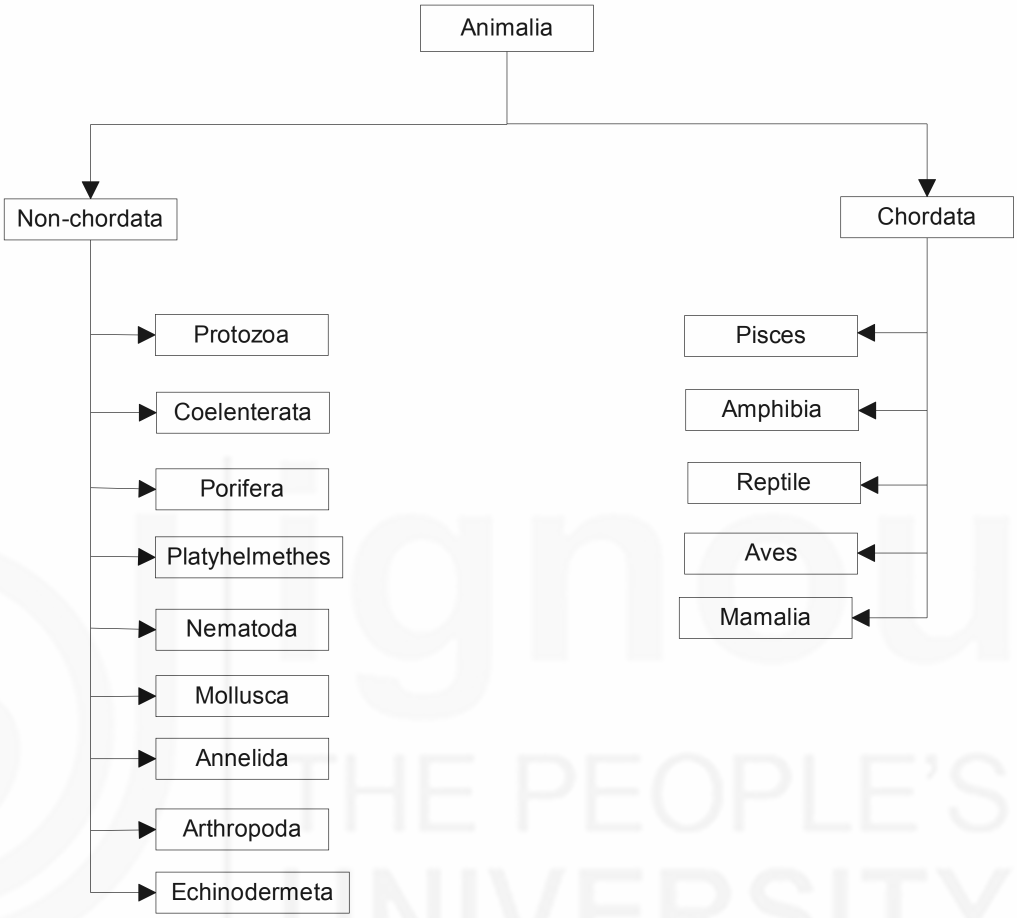


Fig. 14.8: Taxonomy of Animal Kingdom

In Computer Science, Ontologies are expressed in the languages that allow abstraction of concepts. Hence, ontology can be defined as "as a level of abstraction of data models, analogous to hierarchical and relational models, but intended for modeling knowledge about individuals, their attributes, and their relationships to other individuals". (Ref. 10)

14.4.2 Web Ontology

Presently, search engines perform searching over stored indexes in their databases with pattern match algorithm. This search lacks representation of concept with search term. This inherent problem is not due to any difficulty with search engines rather it is due to representation of data in webpage using Hyper Text Markup Language (HTML), the language of the Web. Hence, a mechanism is visualized to represent the data of web pages using another language i.e. Extensible Markup Language (XML) with a standard data description framework called as Resource Description Framework (RDF). It is understood that each individual web page can be considered as an entity and will have its attributes or characteristics. Based on this property the pages can be grouped and further they can form relation with other web page(s) or group of web pages. This develops a kind of web based ontology also known as web ontology for web documents but the original idea of ontology remains same. This framework uses standard vocabularies like Resource Description Framework Schema (RDFS) and Web Ontology Language (OWL) for describing the concepts and their relations with other concepts. The search engines extract the data from the web page and preserve the relation with the data, so that meaningful results can be generated.

14.4.3 Types

Generic Ontologies

Generic ontologies cover large spectrum of knowledge domains. They defines concepts at very broad level. Generic ontology represents broad concepts and their relationships. These ontologies are easy to reuse. Generic ontologies represent class of libraries which can be used with different problem domains and environment. These ontologies are like an umbrella ontology which can be further used for more specific purpose in conjunction with more specific ontology. These ontologies provide a mechanism for interoperability among different related ontologies. However, generic ontologies have following key features:

1) Generic ontologies are created from thesaurus, term dictionary or classification schemes and so on.

2) It provides logical concreteness and suitability for information interchange.

3) These ontologies don’t provide any informational explanation for content used.

4) It is suitable to be used with more than one discipline or domain of knowledge.

5) Normally, this kind of ontology lacks sound principles of development or in other words, they follow popular approach.

Core Ontologies

With regard to ontological content there are two schools of thoughts. One claims that content depends highly on the context and hence any ontology prepared can work and only work with the same content or concept. However, the other school suggests that there are ontologies which follow minimal standard vocabulary. The vocabulary used is from philosophy or cognitive science. Hence, the used vocabulary is domain independent or in other words it is only dependent on philosophy and cognitive science. But the content it represents belongs to specific domain of knowledge. This kind of ontology is known as core ontologies.

The core ontology has been used to reach an agreement on the types of entities (and their relationships) needed in a community of practice. It is being used to dynamically negotiate the intended meaning across a distributed community. It has been used to align, integrate and merge several sources of metadata or ontologies. Hence, it can used to build more than one application or service. It can be adopted as a template for specifying the content in some domain.

Hence, the key features of core ontology can be given as:

- the core ontology specialises a foundational or top-level ontology

- the core ontology has been built through a well-motivated methodology that nonetheless avoids the reuse of a foundational ontology

- the core ontology has “built-in” (but explicit) criteria for well-foundedness.

Domain Specific Ontologies

Ontologies are developed keeping specific objectives in mind like, defining various components, describing specific functionality and so on. Often specific ontologies are required based on a specific knowledge sphere, area, field, region or realm. Such ontologies are known as domain specific ontolgoies. For example, ontology of organisational chart can be considered as domain specific ontology. The following example also demonstrates an ontology of still camera.

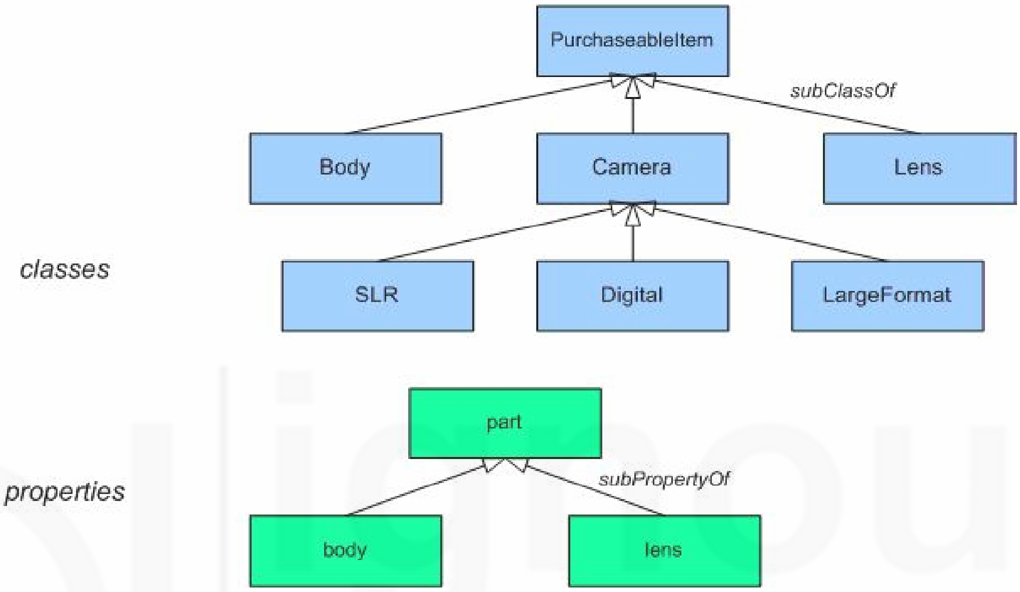


Fig. 14.9: Ontology of Still Camera

Source: http://www.seasr.org/wp-content/plugins/meandre/rdfapi-php/doc/tutorial/img/Camera-classes.png

Key features of Domain Specific Ontologies are as follows:

- It is restricted to a specific domain or area
- It highlights all the components and there interrelations of a domain
- It not only highlights components but also highlights their properties

Task Oriented Ontologies

In a more complex system, the operations are broken into different levels like top level, middle level and inner most level. Each level may have its own objectives as defined by the system analyst. Therefore, each level performs its individual task and transfers the output to next level. A conceptual framework of described system is known as Task Oriented Ontology. Like other ontologies it also consists of taxonomy and axioms. Axioms are rules for reasoning, principles, or constraints among the concepts. Hence, a task oriented ontology has three parts,

- Lexical level
- Conceptual level
- Symbol level

Lexical level

At lexical level task ontology provides human-friendly understanding in terms of which users can easily describe their own task. It provides comprehension for human readability and descriptiveness.

Conceptual level

At conceptual level task ontology simulates the various problem solving processes at the conceptual level and demonstrates the possible solutions through the rules or reasoning. It provides operationality only at conceptual level rather implementation or execution level.

Symbol level

This level provides operationality at implementation or execution level. The ontology makes system run the task description by translating it into instructions.

Self Check Exercises

Note: 1) Write your answers in the space given below.

2) Check your answers with the answers given at the end of this Unit.

4) What is ontology?

....

....

....

....

5) Discuss different parts of task oriented ontology.

....

....

....

....

14.5 INTEROPERABILITY

The term interoperability means working in collaboration. In a distributed service environment, different resources work together to produce a common service or goal. It is very common to understand different modules or services of an object/product. It is not desired to know how each module or service is functioning. But all together should able to work in collaboration to produce one service or product. The individual module or service must have enough common ground so that exchange of individual output can be shared without any error and misunderstanding. This requires standardisation of individual output according to some specifications. The standardisation provides common platform for exchange of communication or services. There are several definitions given for interoperability,

According to National Information Standard Organisation (NISO), "Interoperability is the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality"

14.5.1 Need

In the library parlance, concept of interoperability is used since long. The use of MARC21 bibliographic standard or any other bibliographic standard, in conjunction with ISO 2709 or MARC format or MARC XML format provides facility to exchange bibliographic data among libraries. This exchange can be used in various ways like, generating a single platform based search facility for a number of libraries, reusability of library catalogue and so on.

The most important use of interoperability is seen in telephone industry. Irrespective of operators one can make phone calls or send messages. This is because of adherence to one kind of standard. Similarly, emails can be sent across different service providers because of Simple Mail Transfer Protocol (SMTP).

Another example of interoperability is seen in the field of industry. In medicine it is very important to record the case histories, hence an efficient Electronic Health Record (EHR) systems is required. If different hospitals record case histories under their individual specifications then communication and exchange of case studies among hospitals will become impossible. Hence, it is the need of the hour that medicine comes out with a standard for maintaining and managing EHR systems so that such systems can communicate among themselves.

14.5.2 Interoperability and Web Search

Interoperability has a major role to play in the Web parlance. Web is unorganised and a distributed environment. Information over the Internet is presented in HTML format. In order, to improve upon the search results of search engines, use of metadata schemas is thought of. But soon it was realised that there is a plethora of metadata standards. Though, Dublin Core metadata elements given by World Wide Web Consortium (W3C) evolved as a de facto standard for describing web documents but other standards have also made a mark over the scenario like, MARC21, Government Information Locator Service (GILS), e-Government Metadata Standard (e-GMS), Encoded Archival Description (EAD), Geospatial Metadata (GEO) and so on. It has been continuously observed and felt that search engines should come up with some kind of interoperability model so that cross standard search is possible. The concept is known as federated searching.

Currently, search engines read data from metatag <meta> of an HTML document or an XML file which describes the resource in Resource Description Format (RDF). RDF is used for defining onotologies in order to describe web resources and it has broader scope than the former ones.

14.5.3 Methods for Achieving Interoperability

Mapping/Matching

Mapping is one of the methods used for achieving interoperability. Mapping means relating or corresponding one to one between the entities of two sets. Mappings between two ontologies means establishing correspondence between each entity of ontology A against entities of ontology B with respect to their meanings. The mapping does not lead to creating new set of entities rather it only produces correspondence.

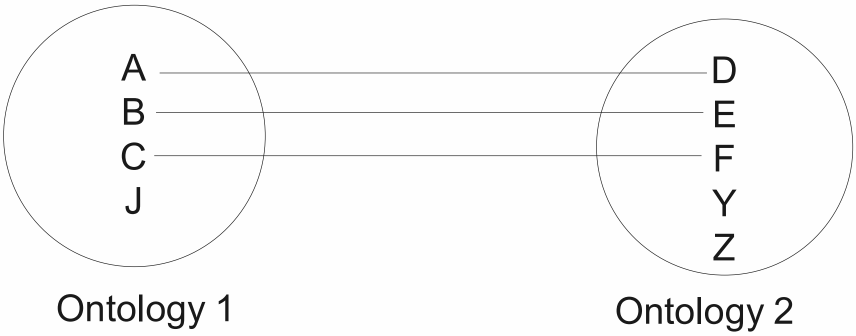


Fig. 14.10: Mapping of Ontologies

Alignment

Ontology alignment is a process of bringing different ontologies into mutual agreement. This process involves bringing ontologies together such that redundancies are removed and logical elements are kept. Hence, the process requires transformation of the involved ontologies. However, any element which is expected in the mutual ontology may also be included. Therefore, alignment might bring a complete new picture of ontology.

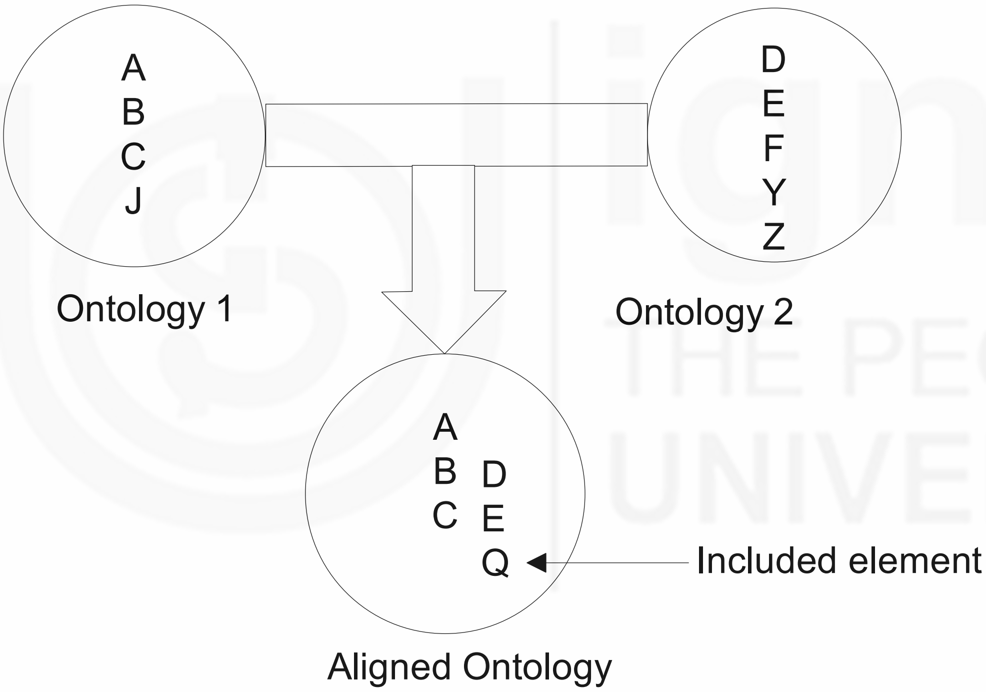


Fig. 14.11: Alignment of Ontology

Transformation

Transformation leads to complete change in the original ontology. The change may occur in terms of elements, attributes or concepts. Hence, the resultant ontology would be a completely new ontology based on the previous one. However, the degree of change in the structure or semantics may vary depending on situation. The process of transformation may be additive or subtractive depending on the needs of the original.

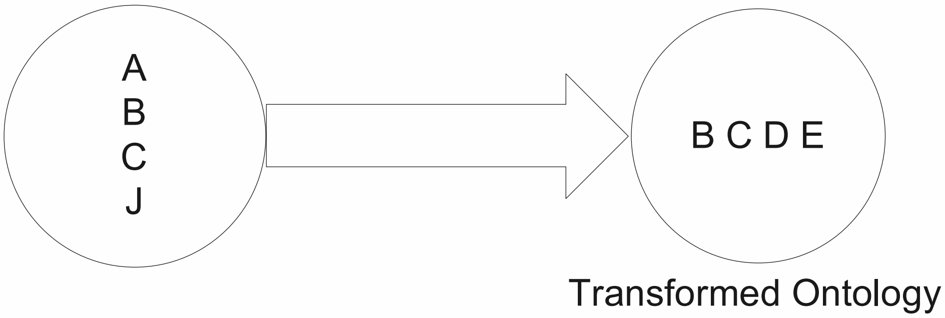


Fig. 14.12: Transformation of Ontology

Translation

Often, it so happens that ontology is to be used in different environments. The change of environment can be subject domain, software or language. In such a situation it is required that the original ontology is to be changed according to the new environment. However, it is expected that the conceptual meaning or semantics of the original will not change and remain as close as possible to the original.

Merging/Integrating

When two or more ontologies are merged together and form a new ontology it is known as merging or integrating of ontologies. This process leads to a formation of completely new ontology based on the previous once.

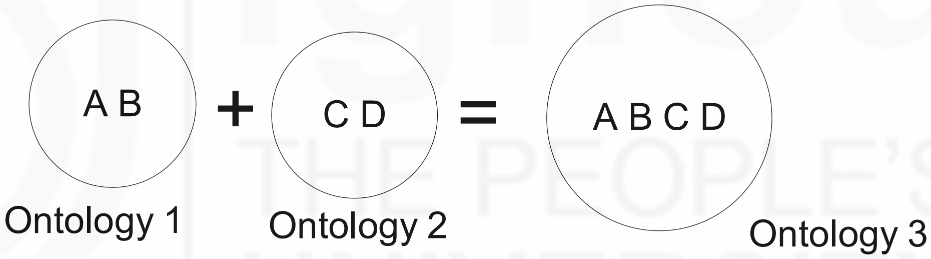


Fig. 14.13: Merging of Ontology

Self Check Exercises

Note: 1) Write your answer in the space given below.

2) Check your answer with the answers given at the end of this Unit.

6) Name different methods of interoperability of metadata.

....

....

....

....

14.5.4 Protocols for Interoperability

The Internet is a source of many online resources containing documents in different formats like, text, graphics audio and video. Individual resources hosting these documents may follow different metadata standards. These documents are to be searched using a search engine. Hence, a cross platform mechanism has been established in the form of protocols to perform searching different resources in one stroke. The whole such system is a distributed system and completely untamed. Developing search agents for such a system is a big challenge. The use of protocols allows users to search several data sources with single effort irrespective of the metadata standard used. Z39.50 (ZEE Thirty Nine point Five Zero), OAI-PMH (Open Archive Initiative and Protocol for Metadata Harvesting) and SRW/U (Search/Retrieve via the Web or URL) are developed for this purpose. Interoperability techniques are still being improved and becoming further sophisticated in order to provide more power and features in the hands of searchers.

There are two protocols which are widely used over the Internet for cross domain search:

- Z39.50 and ZING
- OAI-PMH

Z39.50 AND ZING

The core of interoperable searching is use of protocols. The use of Z39.50 is well accepted and oldest in library services. The protocol was developed to search Online Public Access Catalogues (OPACs) of different libraries. In due course of time, the protocol evolved with several applications like searching deep web (databases over Internet), publishers' catalogue, digital repositories and so on. The protocol performs real-time information retrieval from the source. A Z39.50 server (for example, Zebra server from Index Data www.indexdata.dk) is queried by a Z39.50 client (for example, Yaz client from Index Data www.indexdata.dk). The client searches various Z39.50 servers individually and presents the results of all the servers collectively (refer Fig.14) The server hosting the data must be available in case of using Z39.50 protocol at the time of searching. The only tricky issue in using Z39.50 is mapping of different standards.

The next generation of the Z39.50 protocol is ZiNG (Z39.50-International: Next Generation) maintained by Library of Congress. The protocol is an encapsulation of three protocols

- Z39.50
- Search and Retrieve Web Service (SRW)
- Search and Retrieve URL Service (SRU)

The protocol exploits features of the Z39.50 and web technology. The need of searching multiple domains using WWW created the scope of expansion of Z39.50. The SRU is simple method of searching the Web using GET method and HTTP. The request is carried in name and value pair through URL. The SRW carries a request in the form of a packet known as SOAP (Simple Object Application Protocol). In both the cases result is thrown back in an XML format. The difference between SRW and SRU is that SRW returns XML stream encapsulated in SOAP envelop. The databases are queried with a standard Common Query Language (CQL), a language for database searching. The protocol is supported by several of the search agents specific to libraries. The most important is Library of Congress.

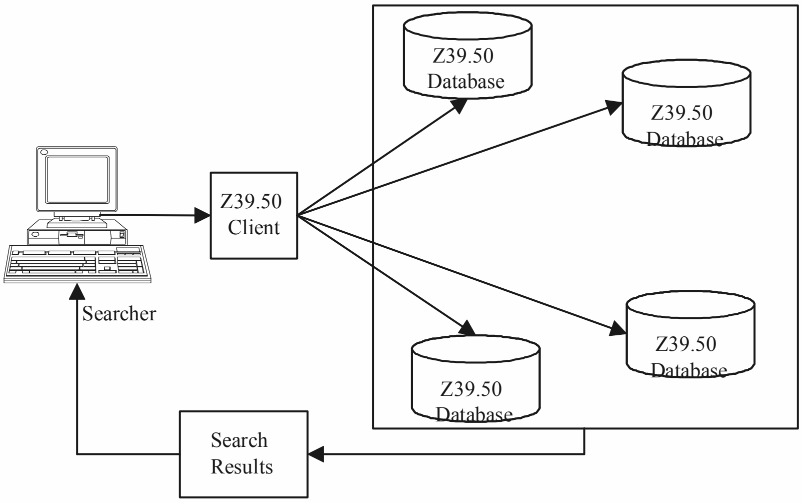


Fig. 14.14 : Model of Basic Z39.50 Protocol

Open Archive Initiative and Protocol for Metadata Harvesting (OAI – PMH)

This is also an HTTP embedded protocol used extensively for interoperable searching and retrieval. The protocol is simple and developed for searching across digital repositories. OAI-PMH intermediary (service provider or search agent) harvests metadata in anticipation from the distributed resources and offers search to the clients. Harvesting means extracting metadata from different resources and storing inside own database. Searching is done by intermediary service provider using its own harvested metadata. However any request for the searched document is directed to the resource or repositories. In case of using OAI-PMH search can also be made to a resource even if that is not available. The result is returned with an XML data format. OAI-PMH supports Dublin core elements.

This protocol has been supported by many digital libraries around the world. OAIster harvester of University of Michigan provides 15,601,208 records from 944 repositories. In India, Search Digital Libraries (SDL) at Documentation Research & Training Centre, Bangalore provides 21220 records from 9 different repositories. This service is first of its kind in India. DP9 at Old Dominion Universtiy is a harvester to enable search engines harvests records from OAI-PMH repositories.

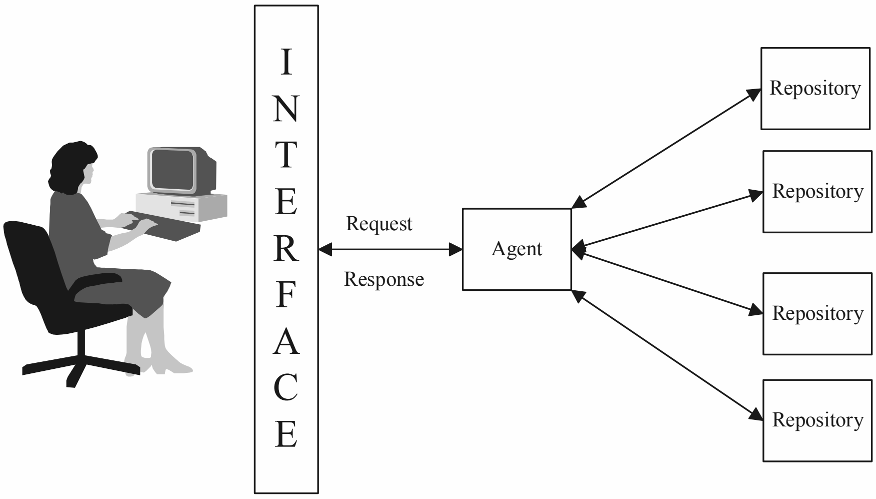


Fig. 14.15: Model of Metadata Harvesting with OAI-PMH

14.6 SUMMARY

The objective of web indexing is to make a website searchable and browsable for the intended users. Web index helps search engines to store more meaningful keywords about the website. Search engines are automated tools for searching the web pages and metadata is one approach towards rendering more semantic based keyword to the search engines. Initially, meta tag <meta> was used for rendering the keyword within a webpage. But it is realised that this approach fails to store the context of search terms. Henceforth, use of ontologies came into being and Resource Description Framework (RDF) is used for representation of intended knowledge with context for automated data extraction by search engines. Sooner it is realized that there is going to be flood of ontologies and metadata schemas which gave rise to the concept of interoperability among the standards.

14.7 ANSWERS TO SELF CHECK EXERCISES

1) Index has a great role to play in the web parlance:

   - Index demonstrates the relationship of topics.
   - Index is used as a source for searching. The search engines use index terms stored inside their database.
   - A good index provides visibility of all the available literature.
   - Index brings material related to what user is looking for, with the help of related terms or concepts.
   - Index focuses on the content of the document and demonstrates the inconsistencies about the treatment of the topics. Helping authors to review the writing for completeness of content.

2) There are different types of Web Indexes:

   a) **Hyperlinked A-Z indexes**

   In web environment, A-Z index is a web page or a group of pages. Each entry in the web page is hyperlinked to a topic or to be more precise to the anchor tag of the resource. The list may also contain synonymous terms linked to the same resource. In an alphabetical index, terms may be written in normal order or in displaying other suitable order.

   b) **Meta-tag keywords indexing**

   Metadata is normally referred as ‘data about some object’. The object could be anything. The data about the object reflects the properties of object for example, Title, Author, Place, Publisher etc. are used to describe a book. Each metadata entry of webpage is used as an index term or if phrase, is further broken in keywords. With such index context of the keywords or phrase is also extracted with the name of meta-tag. This kind of index is known as Meta-tag keywords indexing

   c) **Keyword creation for search engine optimization**

   Search engines look for search term which is queried in its database and fetch the result. If a document appears first in the order of search result then it is said that the page has better visibility. An attempt to get better visibility is known as search engine optimization.

d) **Taxonomies/categories**

Taxonomy refers to abstract structure of subject. It is also referred as subject-based classification. Taxonomy typically displays the hierarchical structure of various components or sub-disciplines. Use of taxonomy for the purpose of indexing, facilitates grouping the like objects or documents together. It displays all the objects or documents which belong to one category. Taxonomy is a kind of Controlled vocabulary. Hence, it can be also used as authority control.

e) **Thesauri**

Thesauri are also a kind of controlled vocabulary. Thesaurus is taxonomy with enhanced functionalities. Thesaurus demonstrates the relation of terms with respect to Broader Terms (BTs), Narrower Term (NTs), Related Terms (RTs), Synonymous Terms (SNs), Usage, Top Term (TT) and so on. The terms in a thesaurus are usually listed alphabetically. (Ref. 2)

f) **Site maps**

Sitemap index is an XML file (Extensible Markup File), which is prepared in a particular format and submitted to search engine. There are programs available over Internet which generates XML based sitemap index. This file can be downloaded and kept in the root directory, when search engine's crawler visits the site it picks up *sitemap.xml* file.

3) There are different kinds of metadata.

There are different types of metadata:

Administrative metadata

When a document is created there are several kinds of information also generated with it. These information are valid and useful during the whole life span of document. These data are stored in as Administrative metadata

Technical Metadata

The technical metadata stores information regarding the file type and associated content type and how it should be rendered. It stores information regarding the how the bytes should be read or in other words how the file should be read. Apart from this it also stores information regarding size or the extent of the file.

Structural Metadata

Structural metadata or structural map of an object explains different components and their role. This handles various sections and sub-sections of the documents and their corresponding relations and roles.

Descriptive Metadata

The metadata used for describing the documents is descriptive metadata. AACR2 and MARC21 are good example of descriptive metadata. Descriptive metadata stores information regarding title, author, place, publisher and so on. This metadata set is important for identification for locating the documents.

Preservation Metadata

One of the most important metadata set used for digital preservation is Preservation metadata. Digital preservation is process of increasing longitivity of documents from physical deterioration. The deterioration of digital objects is against time, technology, media and transfer. In order to secure document and its original features libraries, archive and museum need some kind of documentation in a form of metadata.

4) Ontology studies of existence of entities and their relationships. The relationship is derived due to grouping the entities based on formed groups. These groups are formed due to likeness or similarities of characteristics or attributes of individual entities. The relationship is depicted in form of hierarchy and subdivisions.

5) There are three parts of Task Oriented Ontology:

Lexical level

At lexical level task ontology provides human-friendly understanding in terms of which users can easily describe their own task. It provides comprehension for human readability and descriptiveness.

Conceptual level

At conceptual level task ontology simulates the various problem solving processes at the conceptual level and demonstrates the possible solutions through the rules or reasoning. It provides operationality only at conceptual level rather implementation or execution level.

Symbol level

This level provides operationality at implementation or execution level. The ontology makes system run the task description by translating it into instructions.

6) Different methods of Interoperability of metadata are:

- Mapping/matching
- Alignment
- Transformation
- Translation
- Merging/integrating

14.8 KEYWORDS

Bandwidth : In computer networks, bandwidth is often used as a synonym for data transfer rate - the amount of data that can be carried from one point to another in a given time period (usually a second). This kind of bandwidth is usually expressed in bits (of data) per second (bps).

Browser : A Client program (software) that is used to look at various kinds of Internet resources.

Client : A software program that is used to contact and obtain data from a Server software program on another computer, often from a distance.

Domain Name : The unique name that identifies an Internet site. Domain Names always have 2 or more parts, separated by dots. The part on the left is the most specific, and the part on the right is the most general. For example *ignou.ac.in*

Download : Transferring data (usually a file) from one computer to another computer.

Email : Also known as Electronic Mail, is messages, usually text, sent from one person to another via computer. E-mail can also be sent automatically to a large number of addresses.

Home Page (or Homepage) : Originally, the web page that your browser is set to use when it starts up. The more common meaning refers to the main web page for a business, organisation, person or simply the main page out of a collection of web pages i.e. index page.

Host : Any computer on a network that is a repository for services available to other computers on the network.

HTML (HyperText Markup Language) : The coding language used to create hypertext documents for use on the World Wide Web.

HTTP (HyperText Transfer Protocol) : The protocol for moving hypertext files across the Internet.

Hypertext : Generally, any text that contains links to other documents - words or phrases in the document that can be chosen by a reader and which cause another document to be retrieved and displayed.

Internet : The vast collection of inter-connected networks that are connected using the TCP/IP protocols and that evolved from the ARPANET of the late 60's and early 70's. Also known as 'network of networks'.

Meta Tag : A specific kind of HTML tag that contains information not normally displayed to the user. Meta tags contain information about the page itself, hence the name ("meta" means "about this subject") Typical uses of Meta tags are to include information for search engines to help them better categorize a page.

Network : Any time connecting more than one computer together so that they can share resources, is known as computer network.

Protocol : Protocols are rules define an exact format for communication between computers. For example HTTP protocol defines the format for communication between web browsers and web servers.

RDF (Resource Definition Framework) : The Resource Description Framework (RDF) is a general framework for how to describe any Internet resource such as a Web site and its content. An RDF description (such descriptions are often referred to as metadata, or "data about data") can include the authors of the resource, date of creation or updating, the organisation of the pages on a site (the sitemap), information that describes content in terms of audience or content rating, key words for search engine data collection, subject categories, and so forth.

Search Engine : A (usually web-based) system for searching the information available on the Web.

SEO (Search Engine Optimization) : The practice of designing web pages so that they rank as high as possible in search results from search engines.

Server : A computer, or a software package, that provides a specific kind of service to client software running on other computers.

SMTP (Simple Mail Transfer Protocol) : The main protocol used to send electronic mail from server to server on the Internet.

SOAP (Simple Object Access Protocol) : A protocol for client-server communication that sends and receives information "on top of' HTTP. The data sent and received is in a particular XML format specifically designed for use with SOAP.

SQL (Structured Query Language) : A specialized language for sending queries to databases.

Terminal : A device that allows you to send commands to a computer somewhere else.

URI — (Uniform Resource Identifier) : An address for s resource available on the Internet.

URL — (Uniform Resource Locator) : The term URL is basically synonymous with URI. URI has replaced URL in technical specifications.

URN — (Uniform Resource Name) : A URI that is supposed to be available for a long time. For an address to be a URN some institution is supposed to make a commitment to keep the resource available at that address.

Web page : A document designed for viewing in a web browser. Typically written in HTML. A web site is made of one or more web pages.

Website : The entire collection of web pages and other information (such as images, sound, and video files, etc.) that are made available through what appears to users as a single web server.

XML (eXtensible Markup Language) : A widely used system for defining data formats. XML provides a very rich system to define complex documents and data structures such as invoices, molecular data, news feeds, glossaries, inventory descriptions, real estate properties, etc. As long as a programmer has the XML definition for a collection of data (often called a "schema") then they can create a program to reliably process any data formatted according to those rules. Libraries use XML for bibliographic data exchange.

Z39.50 A NISO and ISO standard protocol for cross-system searchand retrieval. Officially, international standard, ISO 23950, Information Retrieval (Z39.50): Application Service Definition and Protocol Specification, and ANSI/ NISOstandard Z39.50.

14.9 REFERENCES AND FURTHER READING

A-Z Indexes to Enhance Site Searching Web. 24 September 2012. <http://www.digital-web.com/articles/a_z_indexes_site_searching/>

AFS Ethnographic Thesaurus. Web. 24 September 2012. <http://www.afsnet.org/?page=AFSET >

Berners-Lee, T., Hendler, J., and Lassila, O. "The Semantic Web". Scientific American, 284.5 (May 2001) 34-43. Print.

Bray, T. "Measuring the Web". Computer Networks and ISDN Systems. 28(7-11-May), 992-1005, 1996. Print.

Browne, Glenda Michelle. "Indexing Web Sites: A Practical Guide". Internet Reference Quarterly. 5.3 (Sep 2001): 27-41. Print.

Diego Calvanese, Giuseppe De Giacomo, and Maurizio Lenzerini. "Description Logics for Information Integration". Ed. Kakas, A.C and F. Sadri. Computational Logic: Logic Programming and Beyond: Kowalski Festschrift, LNAI 2407. Berlin: Springer Verlag , 2002. 41–60. Print.

Genesereth, M. R. and Nilsson, N. J. Logical Foundations of Artificial Intelligence. San Mateo, California: Morgan Kaufman Publishers, 1987. Print.

Gruber, T., "A Translation Approach to Portable Ontology Specifications". Knowledge Acquisition 5.2 (1993), 199-220. Print.

Gruber, T., "Towards Principles for the Design of Ontologies Used for Knowledge Sharing". International Journal of Human and Computer Studies 43. 5/6 (1995), 907-928. Print.

Gruber, Tom "Ontology". In the Encyclopedia of Database Systems, Ed. Ling Liu and M. Tamer Özsu . Springer-Verlag, 2009. Web. 24 September 2012. <http://tomgruber.org/writing/ontology-definition-2007.htm>

Guarino, N., "Formal Ontology in Information Systems", Proceedings of FOIS'98, Formal Ontology in Information Systems, Trento, 3-15, 1998. Print.

Halpin, T.A. Information Modeling and Relational Databases: From Conceptual Analysis to Logical Design. San Francisco, California: Morgan Kaufman Publishers, 1999.

Hruby, Pavel. Ontology-based Domain-driven Design. Web 24 September 2012. <www.softmetaware.com/oopsla2005/hruby.pdf>

Kohler, Jacob. "Ontology Based Text Indexing and Querying for the Semantic Web". Knowledge-Based Systems 19 (2006) 744–754. Web. 24 September 2012. <http://www.researchgate.net/publication/222566181_Ontology_based_text_indexing_and_querying_for_the_semantic_web>

Pretorius, A. Johannes. Ontologies: Introduction and Overview. Web. 24 September 24 2012.

http://www.starlab.vub.ac.be/teaching/Ontologies_Intr_Overv.pdf

Maedche, A.D. Ontology Learning for the Semantic Web. Norwell, Massachusetts: Kluwer Academic Publishers, 2003. Print.

Meersman, R., "Can Ontology Theory Learn from Database Semantics?" Proceedings of the Dagstuhl Seminar 0121, Semantics for the Web, Dagstuhl, 2000. Print.

Meersman, R., "Ontologies and Databases: More than a Fleeting Resemblance" Ed. D'Atri,A. and Missikoff, M. OES/SEO 2001 Workshop, Rome, 2001. Print.

Mitchell, W.J. City of Bits: Space, Place and the Infobahn. Cambridge, Massachusetts: MIT Press, 1999. Print.

Noy, N.F. and Hafner, C.D. "The State of the Art in Ontology Design – A Survey and Comparative Review". AI Magazine. 36.(1997), 53-74. Print.

Ontologies and Knowledge Base. http://briefs.cs.hut.fi/phase8/Ontologies_and_KB/noframes.html

Ontology (information science) http://en.wikipedia.org/wiki/Ontology_(computer_science)

Ontology Interoperability. Draft version 0.3.2. Web. 24 September 2012. <www.dsi.uniroma1.it/~estrinfo/3.%20Ontology%20merging.pdf>

Pisanelli, D.M., Gangemi, A. and Steve, G., "Ontologies and Information Systems: The Marriage of the Century". Proceedings of the LYEE Workshop, Paris, 2002. Print.

Maislin, Seth A. Indexing Online: The New Face of an Old Art. Web. 24 September 2012. < http://taxonomist.tripod.com/websmarts/onlineindexing.html>

Shahar, Yuval, Miksch, Silvia and Johnson, Peter. A Task-Specific Ontology for the Application and Critiquing of Time-Oriented Clinical Guidelines. Web 24 September. <www.springerlink.com/index/d4483r305758786r.pdf>

Summit on Serials in the Digital Environment. Web. http://www.loc.gov/acq/conser/glossary.html

Understanding Metadata. ISBN 1-880124-62-9 www.niso.org/standards/resources/UnderstandingMetadata.pdf

Web Indexing SIG (Special Interest Group). www.webindexing.org

Wikipedia. http://en.wikipedia.org/wiki/Web_indexing

Taylor, Chris. Introduction to Metadata. Web. 24 September 2012. <http://www.itb.hu/fejlesztesek/meta/hgls/core/Background/An_Introduction_to_Metadata.htm>

Zhao, G. and Meersman, R. "Ontology Modelling for Collaborative Business Processes". StarLab Technical Report. Brussels: Vrije Universiteit Brussel, 2003. Print.