## UNIT 14 OVERVIEW OF WEB INDEXING, METADATA, INTEROPERABILITY AND ONTOLOGIES

**Structure**

<table>
<thead>
<tr>
<th>Section</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.0</td>
<td>Objectives</td>
</tr>
<tr>
<td>14.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>14.2</td>
<td>Web Indexing</td>
</tr>
<tr>
<td>14.2.1</td>
<td>Concept</td>
</tr>
<tr>
<td>14.2.2</td>
<td>Types of Web Indexes</td>
</tr>
<tr>
<td>14.3</td>
<td>Metadata</td>
</tr>
<tr>
<td>14.3.1</td>
<td>Concept</td>
</tr>
<tr>
<td>14.3.2</td>
<td>Types</td>
</tr>
<tr>
<td>14.4</td>
<td>Ontology</td>
</tr>
<tr>
<td>14.4.1</td>
<td>Concept</td>
</tr>
<tr>
<td>14.4.2</td>
<td>Web Ontology</td>
</tr>
<tr>
<td>14.4.3</td>
<td>Types</td>
</tr>
<tr>
<td>14.5</td>
<td>Interoperability</td>
</tr>
<tr>
<td>14.5.1</td>
<td>Need</td>
</tr>
<tr>
<td>14.5.2</td>
<td>Interoperability and Web Search</td>
</tr>
<tr>
<td>14.5.3</td>
<td>Methods for Achieving Interoperability</td>
</tr>
<tr>
<td>14.5.4</td>
<td>Protocols for Interoperability</td>
</tr>
<tr>
<td>14.6</td>
<td>Summary</td>
</tr>
<tr>
<td>14.7</td>
<td>Answers to Self Check Exercises</td>
</tr>
<tr>
<td>14.8</td>
<td>Keywords</td>
</tr>
<tr>
<td>14.9</td>
<td>References and Further Reading</td>
</tr>
</tbody>
</table>

### 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 an 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 keywords 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:
Recent Developments

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.

![Fig. 14.2: Hyperlinked A-Z Index](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:
Recent Developments

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**

<table>
<thead>
<tr>
<th>Search engine</th>
<th>Robot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>Googlebot</td>
</tr>
<tr>
<td>Yahoo</td>
<td>Slurp</td>
</tr>
<tr>
<td>MSN</td>
<td>MSNbot</td>
</tr>
</tbody>
</table>

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.

**Fig. 14.3: Metatag in HTML Document**

**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
Recent Developments

- 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

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

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.

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
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

<table>
<thead>
<tr>
<th>Creator</th>
<th>Aditya Tripathi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publisher</td>
<td>Documentation Research &amp; Training Centre</td>
</tr>
<tr>
<td>Identifier</td>
<td><a href="http://drtc.isibang.ac.in">http://drtc.isibang.ac.in</a></td>
</tr>
<tr>
<td>Subject</td>
<td>Library and Information Science</td>
</tr>
<tr>
<td>Format</td>
<td>txt/html</td>
</tr>
<tr>
<td>Language</td>
<td>English</td>
</tr>
<tr>
<td>Rights</td>
<td>Indian Statistical Institute</td>
</tr>
</tbody>
</table>
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.
Recent Developments

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:

```xml
<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>
  </mets:div>
</mets:structMap>
```

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.

![Taxonomy of Animal Kingdom](image)

**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 define 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 ontologies. For example, ontology of organisational chart can be considered as domain specific ontology. The following example also demonstrates an ontology of still camera.

![Ontology of Still Camera](http://www.seasr.org/wp-content/plugins/meandre/rdfapi-php/doc/tutorial/img/Camera-classes.png)

**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 their 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 ontologies 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.
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.

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.
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 one.

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.
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 University is a harvester to enable search engines harvests records from OAI-PMH repositories.
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.
Recent Developments

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
Recent Developments

**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) 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 a 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.


14.9 REFERENCES AND FURTHER READING


Overview of Web Indexing, Metadata, Interoperability and Ontologies


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


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>


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


